PB-242 579

STUDIES OF POLLUTANT CONCENTRATION FREQUENCY DISTRIBUTIONS

California University

January 1975

DISTRIBUTED BY:

NTIS

National Technical Information Service
U. S. DEPARTMENT OF COMMERCE

YL 163051

EPA-650/4-75-004

January 1975

PB242579

Environmental Monitoring Series

STUDIES OF POLLUTANT CONCENTRATION FREQUENCY DISTRIBUTIONS

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
Springfield, Va. 22151

Meteorology Laboratory
National Environmental Research Center
Office of Research and Development
U.S. Environmental Protection Agency
Research Triangle Park, N.C. 27711

93

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA-650/4-75-004	2. PB 242 579
4. TITLE AND SUBTITLE: STUDIES OF POLLUTANT CONCENTRATION FREQUENCY DISTRIBUTIONS	5. REPORT DATE: January 1975
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S): Richard I. Pollack	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Lawrence Livermore Laboratory, University of California, Livermore, California 94550	10. PROGRAM ELEMENT NO. 1AA009
	11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS: Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, N.C. 27711	13. TYPE OF REPORT AND PERIOD COVERED: Final
	14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

16. ABSTRACT: Early air pollution research focused on determining the identity of the concentration distributions for a variety of pollutants and locations and the relationships between attributes of the data, e.g. mean values, maximum levels and averaging times, from an empirical standpoint. This report attempts to identify the nature of the frequency distributions for both reactive and inert pollutants, for both point and area sources, and to some extent for different types of atmospheric conditions using a substantially non-empirical approach. As an illustration of the applicability of these results, a predictive model and a monitoring scheme are proposed based upon knowledge developed by studying the frequency distributions.

It is found that a theory of the genesis of pollutant concentrations based upon the Fickian diffusion equation predicts that concentration distributions due to area sources will be approximately lognormal over a diurnal cycle in the absence of nearby strong sources. It is determined that reactive pollutants will have larger standard geometric deviations than relatively inert pollutants. Empirical observations are in good agreement with these results. The frequency distribution of the logarithms of concentrations due to point sources is derived and shown to be a sum of normal and chi-squared components, with the identity of the dominant term determined by meteorological conditions. This result provides a framework for resolving apparently conflicting results in the literature. The lognormality of other meteorological variables, notably windspeeds and the rate of energy dissipation in turbulent flow, and their relation to air quality frequency distributions is discussed. There is considerable discussion in the literature concerning whether the lognormal distribution provides the best fit. Other distributions that fit air quality data fairly well are investigated, and their mathematical similarity to the lognormal is demonstrated.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Air pollutants Frequency distribution Monitoring Modeling		

18. DISTRIBUTION STATEMENT: Unlimited	19. SECURITY CLASS *(This Report)*: Unclassified	21. NO. OF PAGES: 92
	20. SECURITY CLASS *(This page)*: Unclassified	22. PRICE: 5.45

EPA Form 2220-1 (9-73)

U.S. GOVERNMENT PRINTING OFFICE: 1975 - 640-881/659 - Region 4

EPA-650/4-75-004

STUDIES OF POLLUTANT CONCENTRATION FREQUENCY DISTRIBUTIONS

by

Richard I. Pollack

Lawrence Livermore Laboratory
University of California
Livermore, California 94550

Program Element No. 1AA009

National Environmental Research Center
Office of Research and Development
U.S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

January 1975

RESEARCH REPORTING SERIES

Research reports of the Office of Research and Development, U. S. Environmental Protection Agency, have been grouped into series. These broad categories were established to facilitate further development and application of environmental technology. Elimination of traditional grouping was consciously planned to foster technology transfer and maximum interface in related fields. These series are:

1. ENVIRONMENTAL HEALTH EFFECTS RESEARCH
2. ENVIRONMENTAL PROTECTION TECHNOLOGY
3. ECOLOGICAL RESEARCH
4. ENVIRONMENTAL MONITORING
5. SOCIOECONOMIC ENVIRONMENTAL STUDIES
6. SCIENTIFIC AND TECHNICAL ASSESSMENT REPORTS
9. MISCELLANEOUS

This report has been assigned to the ENVIRONMENTAL MONITORING series. This series describes research conducted to develop new or improved methods and instrumentation for the identification and quantification of environmental pollutants at the lowest conceivably significant concentrations. It also includes studies to determine the ambient concentrations of pollutants in the environment and/or the variance of pollutants as a function of time or meteorological factors.

Copies of this report are available free of charge to Federal employees, current contractors and grantees, and nonprofit organizations - as supplies permit - from the Air Pollution Technical Information Center, Environmental Protection Agency, Research Triangle Park, North Carolina 27711. This document is also available to the public for sale through the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402.

EPA REVIEW NOTICE

This report has been reviewed by the Office of Research and Development, Environmental Protection Agency, and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the Environmental Proctection Agency, nor does mention of trade names constitute endorsement or recommendation for use.

Publication No. EPA-650/4-75-004

PREFACE

Air quality data have been analyzed as a function of frequency, maxima, the form of the frequency distribution, and averaging time in the 15 papers composing the "Proceedings of the Symposium on Statistical Aspects of Air Quality Data" (U.S. Environmental Protection Agency Report No. EPA-650/4-74-038, Research Triangle Park, North Carolina 27711, October 1974).

Dr. Pollack has drawn on these and other data in his analysis of air pollutant concentration data and the frequency distributions used to describe such data. His dissertation identifies the nature of the frequency distributions for both reactive and inert pollutants, for both point and area sources, and to some extent for different types of atmospheric conditions, using a substantially non-empirical approach. Because of the valuable information presented in his dissertation, Dr. Pollack and the Lawrence Livermore Laboratory have given their kind permission to the Meteorology Laboratory of EPA to publish it for wider distribution.

TABLE OF CONTENTS

		Page
I	INTRODUCTION	1
	The Problem	1
	The Research	2
	The Significance	3
II	AIR QUALITY MEASUREMENTS	5
	The Derivation of a Frequency Distribution of a Pollutant Emitted from a Point Source	5
	Area Source	6
	An Extension of the "Gaussian Plume" Point Source Distribution Derivation	7
	A New Approach to the Derivation of Frequency Distributions of Pollutant Concentrations	8
	Lognormality Over Various Averaging Times	12
	Summary of Derivations	18
	Conclusion	19
III	FREQUENCY DISTRIBUTIONS OF RELATED VARIABLES	20
	Advection	20
	Diffusion	21
	Particle Sizes	25
	Conclusion	26
IV	FREQUENCY DISTRIBUTIONS FOR VARIOUS POLLUTANTS AND SOURCE TYPES	27
	Reactive versus Inert Pollutants	27
	Point versus Area Sources	37
	Summary	38
V	THE FREQUENCY DISTRIBUTIONS	39
	Lognormal Distribution	39
	Weibull Distribution	49
	Gamma Distribution	52
	Pearson Distribution	56
	Mathematical Similarity	58
	Summary	61

Preceding page blank

		Page
VI	ILLUSTRATIVE APPLICATIONS	62
	Analysis of Meteorological Patterns for Pollution Level Forecasting	62
	Selecting the Clusters	66
	Classifying New Days	66
	Recalibration	67
	Spatial Interpolation	68
	An Example	68
	Development	72
	Transition Matrices	72
	Random Sampling	73
VII	SUMMARY AND CONCLUSIONS	75
	Area Sources	75
	Point Sources	76
	Related Variables	77
	Other Distributions	77
	Applications	78
	Future Research	78
	LITERATURE CITED	80

TABLE OF FIGURES AND GRAPHS

		Page
Fig. 1(a).	Autocorrelation versus lag for CO concentrations in San Francisco, 1970 hourly averages.	14
Fig. 1(b).	Autocorrelation versus lag for CO concentrations in San Francisco, 1970 hourly averages (log plot).	15
Fig. 1(c).	Autocorrelation at lag 1 versus averaging time for CO concentrations in San Francisco, 1970 hourly averages.	16
Fig. 2.	Concentration and windspeed frequency distributions for CO and windspeed for San Francisco, 1970 hourly averages.	22
Fig. 3.	Probability distribution of the squared temperature difference compared with lognormality. $P(\epsilon < \epsilon_0)$. $\epsilon = (\Delta T)^2/[(\Delta T)^2]$ (15). Separation = 2 cm, 10^4 samples per plot.	24
Fig. 4(a).	Oxidant concentration versus time in Los Angeles, hourly averages.	28
Fig. 4(b).	CO concentrations versus time in San Francisco, hourly averages.	29
Fig. 5(a).	SO_2 concentrations, direction 289°-308°, Lacq, France, 1968-1969 (3).	31
Fig. 5(b).	NO_2 concentrations, direction 108°-121°, Lacq, France, 1968-1969 (3).	32
Fig. 5(c).	SO_2 concentrations, direction 108°-121°, Lacq, France, 1968-1969 (3).	33
Fig. 5(d).	NO_2 concentrations, direction 289°-308°, Lacq, France, 1968-1969 (3).	34
Fig. 6(a).	Log probability plot of oxidant concentrations in Los Angeles, 11/11/70, hourly averages.	35
Fig. 6(b).	Oxidant concentrations in Riverside, California and Los Angeles, California, 1967 hourly averages.	36
Fig. 7.	Frequency curves of the normal and lognormal distributions.	41
Fig. 8.	Frequency curves of the lognormal distribution for three values of σ^2.	42

		Page
Fig. 9.	Frequency curves of the lognormal distribution for three values of μ.	43
Fig. 10(a).	Regions of convergence where the sum of n lognormal variates is approximately lognormal. (A) Convergence for both normal and lognormal approximations, (B) convergence for the lognormal approximation, (C) convergence uncertain (21).	45
Fig. 10(b).	CO concentration in San Francisco, hourly averages.	46
Fig. 10(c).	CO concentration for various categories of pollution days in San Francisco, 1970 hourly averages.	47
Fig. 11.	Frequency curves for Weibull (top) and Rayleigh probability distributions.	51
Fig. 12.	Cumulative Weibull distribution plotted on log probability paper.	52
Fig. 13(a).	Frequency curves for gamma probability distribution for various values of α.	54
Fig. 13(b).	Cumulative gamma distributions plotted on log probability paper.	55
Fig. 14.	Skewness-kurtosis plane in Pearson's system.	57
Fig. 15.	Cumulative beta distribution plotted on log probability paper.	59
Fig. 16.	A possible set of air quality patterns.	63
Fig. 17.	Geometric mean versus standard geometric deviation for individual days for oxidant concentration in Los Angeles, California, 1970 hourly averages.	65
Fig. 18.	A set of clusters of air quality day-types from the data in Fig. 17.	69
Fig. 19.	An example of the form of the chart to be developed in comparing the clusters generated in Fig. 18 with windspeed and temperature.	71

ABSTRACT

Early air pollution research focused on determining the identity of the concentration distributions for a variety of pollutants and locations and the relationships between attributes of the data, e.g.,mean values, maximum levels and averaging times, from an empirical standpoint. This report attempts to identify the nature of the frequency distributions for both reactive and inert pollutants, for both point and area sources, and to some extent for different types of atmospheric conditions using a substantially non-empirical approach. As an illustration of the applicability of these results, a predictive model and a monitoring scheme are proposed based upon knowledge developed by studying the frequency distributions.

It is found that a theory of the genesis of pollutant concentrations based upon the Fickian diffusion equation predicts that concentration distributions due to area sources will be approximately lognormal over a diurnal cycle in the absence of nearby strong sources. It is determined that reactive pollutants will have larger standard geometric deviations than relatively inert pollutants. Empirical observations are in good agreement with these results. The frequency distribution of the logarithms of concentrations due to point sources is derived and shown to be a sum of normal and chi-squared components, with the identity of the dominant term determined by meteorological conditions. This result provides a framework for resolving apparently conflicting results in the literature. The lognormality of other meteorological variables, notably windspeeds and the rate of energy dissipation in turbulent flow, and their relation to air quality frequency distributions is discussed. There is considerable discussion in the literature concerning whether the lognormal distribution provides the best fit. Other distributions that fit air quality data fairly well are investigated, and their mathematical similarity to the lognormal is demonstrated.

As an illustration of the significance of the results developed herein, a predictive scheme that uses concentration frequency distributions as a basis for classifying meteorological patterns is presented. This scheme uses natural clustering of the distribution parameters to identify meteorological and emission patterns. Finally, an air quality monitoring random sampling scheme based upon the distributions identified in the literature and this work is presented and its improvement over non-parametric techniques is demonstrated.

CHAPTER I—INTRODUCTION

The Problem

In recent years public interest in the quality of ambient air has increased. As information concerning deleterious health effects has gained wider acceptance, government has moved to specify standards for the quality of ambient air. The standards are most often given in terms of a maximum value which may be exceeded only once a year for concentrations averaged over a specified period of time.

The approach taken in relating air quality data to standards is to calculate the frequency distribution for the air quality data, from which concentrations at various averaging times can be derived. It is essential that these distributions be very precise because of the stunning economic impact on a region which must change its way of life to conform to air quality standards. As a result, considerable attention is being paid to this problem.

The earliest work on this problem consisted of the empirical identification of the frequency distributions of surface air pollutant concentrations. Various distributions were proposed with different degrees of success. The most widely accepted of these distributions is the lognormal, primarily due to the work of Larsen (1) who presented data indicating that concentrations of all pollutants in all tested cities for all averaging times are approximately lognormally distributed. It was also noted from these data, however, that some pollutants tended to fit better than others, there were differences between cities, and it was not clear why averages of lognormal variables should be lognormal rather than normal as the Central Limit Theorem would indicate.

Later work determined that different distributions and/or different ranges of parameter values were appropriate in different circumstances. Marked differences were noted between inert and reactive pollutants and point and area sources. Conflicting results are presented in the literature concerning these distributions. It is clear that an understanding of these results is important because of the economic impact of the decision that ambient air quality standards

(AAQS) are being violated. Further, an understanding of the nature of the distributions and their parameters in the various cases can serve only to enhance our understanding of the fundamental principles involved.

There are, of course, more pragmatic applications of this research. In particular, the formulation and validation of air pollution models cannot proceed without some knowledge of the form of the output to be expected. The effects of various types of sources on ambient air quality can be estimated through knowledge of the form of the resulting concentration distributions and how the parameters vary with source type, pollutant type and distance from the source. Considerable savings in time and money can be made in air quality monitoring, prediction and modeling through applications of the techniques presented herein.

At present, the scientific community has not reached a concensus concerning the points raised above. There are a number of conflicting empirical results, and there is little work on a more theoretical level. The present work seeks to add new information to the discussion, derived from a non-empirical viewpoint.

The Research

The objective of this work is to present a model which binds together previous theoretical and empirical findings within a unified framework.

First, the frequency distribution of surface air pollutant concentrations is derived starting from the differential equation describing the time evolution of air pollutants through the atmosphere. It is shown that for certain fairly general conditions, the distribution is lognormal.

Using the "Gaussian Plume" equation, which describes the dispersion of a pollutant from a point source as a spatial bivariate normal distribution, the concentration distribution resulting from a point source is derived. It is shown here that the identity of the distribution is dependent upon the distance from the source, the atmospheric stability conditions, and the magnitude of the windspeeds. Within this framework several apparently conflicting results from the literature (2), (3) can be reconciled.

One of the most significant of Larsen's empirical results was the fact that pollutant concentrations are approximately lognormally distributed for a wide spectrum of averaging times. This appears to contradict the central limit theorem of mathematical statistics. However, through the model presented herein the averaging process can be seen as a filter of various scales of atmospheric motion, each of which results in the lognormal.

Several investigators have proposed other distributions for describing air quality data. The Weibull, beta, and gamma distributions are the most often suggested. An obvious question concerns the reason, in mathematical terms, that these distributions fit the same data fairly well. It is presented later that a transformation can be found which reduces the lognormal, gamma, and Weibull distributions to very similar forms. This suggests that there is little significant difference between these distributions for the parameter values often observed.

A variety of other meteorological variables are approximately lognormally distributed, particularly those describing atmospheric motion, or motion of substances suspended in the atmosphere. The relationship between several of these variables and pollutant concentrations is discussed.

Rather than merely stating these derived results, considerable attention has been paid to empirical evidence both from the literature and compiled for this work. The assertions made herein are supported by data which are presented concomitant with the non-empirical results.

The Significance

This study treats the problem of the identification of the frequency distributions of air quality data comprehensively. The following are discussed:

1. Which parametric distributions are appropriate to characterize pollutant concentrations from point and area sources and for inert and reactive pollutants? Why?
2. What is the effect of averaging time on these frequency distributions?

3. How are these distributions affected by other meteorological variables?
4. How can this information be applied?

It is concluded that the information developed here can be applied to developing models for alert level forecasting, air quality monitoring, and more.

CHAPTER II—AIR QUALITY MEASUREMENTS

Air quality data are continuously monitored, with the receptors punching out one reading every 5 minutes. These raw data are then averaged as follows:

let

$$X_1, X_2 \ldots X_i \ldots X_n$$

be a sequence of 5-minute observations. The averages

$$\frac{X_1 + X_2 + \ldots X_k}{K}, \frac{X_{k+1} + X_{k+2} + \ldots X_{2k}}{K}, \frac{X_{2k} + X_{2k+1} + \ldots X_{3k}}{K}$$

are then calculated and referred to as the averages of time K. These averages are the standard concentration measurements used by researchers and analysts.

Air quality standards are set based upon the relationship between air pollution exposure and health effects. A comprehensive exposition of these standards can be found in Ref. (4). Updated information is published by the Environmental Protection Agency (EPA) Office of Air Programs.

To compare ambient air quality to standards, frequency distributions of the averages are calculated. The cumulants of these distributions are used to determine the probability of exceeding a particular standard.

This chapter concerns itself with determining the nature of these distributions from a non-empirical standpoint.

The Derivation of a Frequency Distribution of a Pollutant Emitted from a Point Source

The earliest discussion of pollutant concentration frequency distributions was by Frank Gifford in 1958 (5). Gifford started with the equation describing the diffusion of a plume of stack effluent. This "Gaussian Plume" equation,

$$\frac{X}{R} = (2\pi\bar{Y}^2U)^{-1} \exp\left[-\frac{(y - D_y)^2 + (z - D_z)^2}{2Y^2}\right] \qquad \text{(II-1)}$$

where

R is the continuous rate of emission,
$\bar{Y}^2$ is the variance of the material in individual disk elements,
$\frac{X}{R}$ is the instantaneous relative concentration,
U is the magnitude of the wind vector, and
D_y, D_z are the distances of the plume center from the origin,

can be simplified by defining

$$\mathcal{Y} = \frac{(y - D_y)}{(2\bar{Y}^2)^{1/2}} \qquad \mathcal{Z} = \frac{(z - D_z)}{(2\bar{Y}^2)^{1/2}} \; . \qquad \text{(II-2)}$$

Therefore,

$$\ell = \mathcal{Y}^2 + \mathcal{Z}^2 = -\ln\left(C_1 \frac{X}{R}\right) \qquad \text{(II-3)}$$

where

$$C_1 = 2\pi\bar{Y}^2U \qquad \text{(II-4)}$$

The terms $\mathcal{Y}^2$ and $\mathcal{Z}^2$ are each chi-squared variables, and by the reproductive property of chi-squared variates the result of the convolution is also chi-squared. Hence the natural log of the concentration is directly proportional to a chi-squared random variable.

Note that this result applied to a point source only.

Area Source

In 1972 Gifford expanded this derivation to include area sources by summing a number of point sources (6).

For n sources which affect concentration at the same point, Eq. (II-3) is summed over all n sources yielding:

$$-L = \ln \prod_{i=1}^{n} C_1 \, X_i/Q_i = \sum_{i=1}^{P} \left(Y_i^2 + Z_i^2\right) \qquad \text{(II-5)}$$

which can be written

$$\ln \prod_{i=1}^{n} \left[C_i \, X_i/Q_i\right]^{1/n} = -L/n \qquad \text{(II-6)}$$

where the term within the brackets is the weighted geometric mean and the right-hand side is normally distributed by the Central Limit Theorem. If the geometric and arithmetic means are simply related, e.g. proportional, then the natural logarithm of concentration is normally distributed.

An Extension of the "Gaussian Plume" Point Source Distribution Derivation

The standard Gaussian Plume equation uses U, the mean windspeed, as a parameter. However, there is considerable evidence to the effect that windspeeds are approximately lognormally distributed, and this may be expected to affect the concentration distribution. If we return to

$$\ell = \mathscr{Y}^2 + \mathscr{Z}^2 = -\ell n \, C_1 \frac{\chi}{Q} \qquad \text{(II-7)}$$

where

$$C_1 = 2\pi \bar{Y}^2 U \, , \qquad \text{(II-8)}$$

$\mathscr{Y}$ and $\mathscr{Z}$ are defined in terms of the parameters of the Gaussian Plume equation and are assumed normally distributed.

This equation can be written

$$\mathscr{Y}^2 + \mathscr{Z}^2 = -\left[\ell n \, (2\pi \bar{Y}^2 U) + \ell n \left(\frac{\chi}{Q}\right)\right]. \qquad \text{(II-9)}$$

Defining $K_1 = 2\pi\mathcal{V}^2$, one finds

$$\mathcal{Y}^2 + \mathcal{Z}^2 = -\left[\ell n\, K_1 + \ell n\, U + \ell n\left(\frac{\chi}{Q}\right)\right]. \qquad \text{(II-10)}$$

The term $\mathcal{Y}^2 + \mathcal{Z}^2$ is expcnentically distributed or chi-squared with 2df, $\ell n\, K_1$ is a constant, and $\ell n\, U$ is normally distributed. If

$$\mathcal{Y}^2 + \mathcal{Z}^2 \ll \ell n\left(\frac{1}{U}\right) \qquad \text{(II-11)}$$

then the lognormal distribution results. ! Thus the two results are reconciled.

Equation (II-10) is intuitively reasonable for periods of non-negligible wind because $\ell n\left(\frac{1}{U}\right)$ is a measure of advective flux while $\mathcal{Y}^2 + \mathcal{Z}^2$ is a measure of diffusive flux. Except in periods of extremely low winds, the advective flux will be the larger. During low wind periods, apparently the lognormal approximation will be poor. The author knows of no such empirical analysis or theoretical analysis at this time. Indeed, atmospheric modeling during calm conditions is in its infant stage.

In short, the exponential distribution is appropriate under the assumption of constant wind velocity. However, windspeeds are ordinarily lognormally distributed, and the advective flux is of greater significance than the diffusive flux. Under these more general conditions we see that χ/Q is approximately lognormally distributed for periods of non-negligible wind. No conclusion is reached for calm periods save that the lognormal approximation is likely to be poor. We realize, of course, that the concentration must be nonzero at all times for the lognormal to ›e correct. For most pollutants the background concentration is enough to satisfy this condition.

A New Approach to the Derivation of Frequency Distributions of Pollutant Concentrations

Also in 1972, Knox and Pollack (7) derived the following relationship from theoretical considerations, using a substantially different approach.

Consider a stochastic process of the form:

$$X_i = X_{i-1} + X_{i-1} Y_i \tag{II-12}$$

where Y_i is an independent stochastic variable, arbitrarily distributed.

If we solve Eq. (II-12) for Y_i

$$\frac{X_i - X_{i-1}}{X_{i-1}} = Y_i \tag{II-13}$$

and sum both sides

$$\sum_0^N \frac{X_i - X_{i-1}}{X_{i-1}} + \sum_{\ell=0}^N Y_i, \tag{II-14}$$

We can approximate the left side by

$$\int_{n=0}^N \frac{dx}{X} = \sum_{\ell=0}^N Y_i \tag{II-15}$$

$$\ell n\, X_N/X_0 = \sum_{\ell=0}^N Y_i \,. \tag{II-16}$$

By the Central Limit Theorem, $\sum_{\ell=0}^N Y_i$ is normally distributed, hence X_n is lognormally distributed.

This is known as the law of proportional effect; the percentage change in a variable is equal to a constant plus an error. If the absolute change had been equal to this same constant-plus-error term, the normal distribution would have resulted. Hence the lognormal distribution is the result of a multiplicative process, whereas the normal distribution results from an additive process.

If we examine the differential equation describing the time evolution of pollutants in the atmosphere:

$$\frac{d\psi_a}{dt} + u\frac{\partial\psi_a}{\partial x} + v\frac{\partial\psi_a}{\partial y} + w\frac{\partial\psi_a}{\partial z} = \frac{\partial}{\partial x}\left(K_x\frac{\partial\psi_a}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y\frac{\partial\psi_a}{\partial y}\right)$$

$$+ \frac{\partial}{\partial y}\left(K_z\frac{\partial\psi_a}{\partial z}\right) + \frac{S_a(x,y,z,t)}{V} + \frac{P}{V}(\psi_a, \psi_b \ldots \psi_n, t), \qquad \text{(II-17)}$$

where ψ_a is the concentration of pollutant a; u, w, and v are the velocity components; K_x, K_y are the lateral vertical eddy diffusivities which are lognormally distributed based upon the lognormality of ϵ and the reproductive properties; K_z is the vertical eddy diffusivity; S_a is the source term for pollutant a; P is the term representing changes in concentration due to photochemistry; and V is the volume of air for which S and P act.

This equation can be manipulated to represent a box model formulation (8) where we are concerned with the concentration averaged over a box which is surrounded by M other boxes.

$$\frac{d\left[\psi_k(m,t)\right]}{dt} = -\sum_{j=0}^{M}\left[T_A(m,j) + T_D(m,j)\right]\ \psi_k(m,t)$$

$$+ \sum_{j=0}^{M}\left[T_A(j,m) + T_D(j,m)\right]\psi_k(j,t) + S_k(m,t)$$

$$+ P_k\left[\psi_a(m,t) \ldots \psi_n(m,t), t\right] \quad . \qquad \text{(II-18)}$$

Where $T_A(m,j)$ and $T_D(m,j)$ are the advective and eddy diffusive transfer coefficients from box m to box j. The lognormal distribution can be argued for these latter variables in a similar manner as for K_x and K_y.

This equation is also consistent with the generating process, Eq. (II-12), when certain reasonable restrictions hold.

1. The contribution of advection and diffusion terms are larger than the contribution of the source term. It has been found empirically that if this is not the case, lognormality does not result (9).

2. The concentrations in the surrounding boxes are on the average over long periods of time close to that of the box we are interested in because they are subjected to similar stimuli.

These restrictions transform Eq. (II-18) to:

$$\frac{d\psi(m,t)}{dt} = -\sum_{j=0}^{M}\left[T_A(m,j) + T_D(m,j)\right]\ \psi_k(m,t) + \sum_{j=0}^{M}\left[T_A(j,m) + T_D(j,m)\right]\ \psi_k(j,t)\ . \qquad \text{(II-19)}$$

Suppose we let

$$\psi_k(j,t) = \psi_k(m,t) + E_k(j,t)\ , \qquad \text{(II-20)}$$

the equation becomes

$$\frac{d\psi(m,t)}{dt} = -\sum_{j=0}^{M}\left[T_A(m,j) + T_D(m,j)\right]\ \psi_k(m,t) + \sum_{j=0}^{M}\left[T_A(j,m) + T_D(j,m)\right]\ \psi_k(m,t) + \sum_{j=0}^{M}\left[T_A(j,m) + T_D(j,m)\right]\ E_k(j,t)\ . \qquad \text{(II-21)}$$

When we sum both sides to show lognormality, we have for the third term,

$$\sum_{T_0}^{T_R}\frac{\sum_{j=0}^{M}\left[T_A(j,m) + T_D(j,m)\right]\ E_k(j,t)}{\psi(t-\Delta t)}\ . \qquad \text{(II-22)}$$

From meteorological reasoning we note that if the flux term is large, indicating strong winds, the difference between $\psi(j,t)$ and $\psi(m,t)$ will be small. Hence the term tends to zero. Conversely, in the case

where the error term $E_k(j, t)$ is large the flux term will usually be small, indicating light winds. Furthermore, in either case or any combination of cases occurring between T_0 and T_R, we can expect that the sign of the term will vary over a diurnal, weekly or seasonal cycle, implying that the positive and negative terms will cancel each other.

This argument implies that Eq. (II-21) is essentially equivalent to Eq. (II-12), which is consistent with the law of proportional effect.

The solution will be source-dominated only when the magnitude of the source terms is comparable to the magnitude of the current concentration. There is reason to believe (9) that in such cases the concentrations will not be lognormally distributed, as the model indicates. This result has also been noted in investigations of particle size distributions (10).

This reasoning is most easily justified for a well mixed urban region. It is not clear that the lognormal distribution will fit as well for non-urban, poorly mixed areas. We do feel, however, that the characteristics of an area's topography and typical meteorology would have to be highly unusual for (II-22) to be so large that the lognormal distribution would fit poorly.

We have not yet discussed the Δt interval necessary for these results. We recognize that it must be sufficiently small not to obscure the generating process. If, for an extreme example, Δt was six months we would not see the effect of Eq. (II-12) because the effect of ψ_{i-1} on ψ_i would have long since died out. Larsen's (1) data are for 5-minute instantaneous readings. We accept this as an appropriate time scale for our purposes, based on the fact that meteorology certainly does not change enough in a 5-minute period to obscure the relevant correlations.

Lognormality Over Various Averaging Times

When the data are averaged over other time periods within the realm of atmospheric motion, the averaging time acts as a filter which smooths out motions of a smaller time scale. This has the effect of allowing us to see only motion of a time scale comparable to the averaging time in the averaged data. Hence the process described by Eq. (II-21) still holds for larger averaging times, but the T_A, T_D terms

now represent motion of a larger scale. This results in lognormality over a large spectrum of averaging times.

An essentially equivalent relation to Eq. (II-12) is

$$x_i = x_{i-1}\,\epsilon$$

This equation can be transformed to represent a first-order autoregressive stochastic process by taking the logs of both sides to yield:

$$\ln x_i = \ln x_{i-1} + \ln \epsilon.$$

This stochastic process is identified by examining the autocorrelation function to verify that it decays exponentially, and by calculating the partial correlation function to verify that it cuts off after lag 1. The partial correlation coefficient can be thought of as a measure of the independent predictive capabilities of x_{i-n}, without regard to information which is "passed through" x_{i-n+k}.

To verify the hypothesis presented above, these statistics were calculated for natural logs of CO concentrations in San Francisco for 1970 as well as the untransformed observations. The autocorrelation function for lags 1 to 12 appears in Fig. 1(a) and Table 1, and is plotted on a log scale in Fig. 1(b). The agreement with the exponential curve appears good. The partial correlation coefficient is equal to the autocorrelation coefficient for lag 1, of course, but thereafter it is negligible statistically. In particular, the values are 0.874, -0.046, and 0.015 respectively for the first three lags. The additive model does not appear to fit the first-order autoregressive model as well.

Figure 1(c) indicates that the multiplicative model appears to be more appropriate for averaging times of 1 hour to 160 hours. A statistical test on the differences between the autocorrelations from the multiplicative as opposed to the additive model was performed. The results indicated that the autocorrelations are indeed significantly different. Table 2 presents several such calculations for representative lags and averaging times.

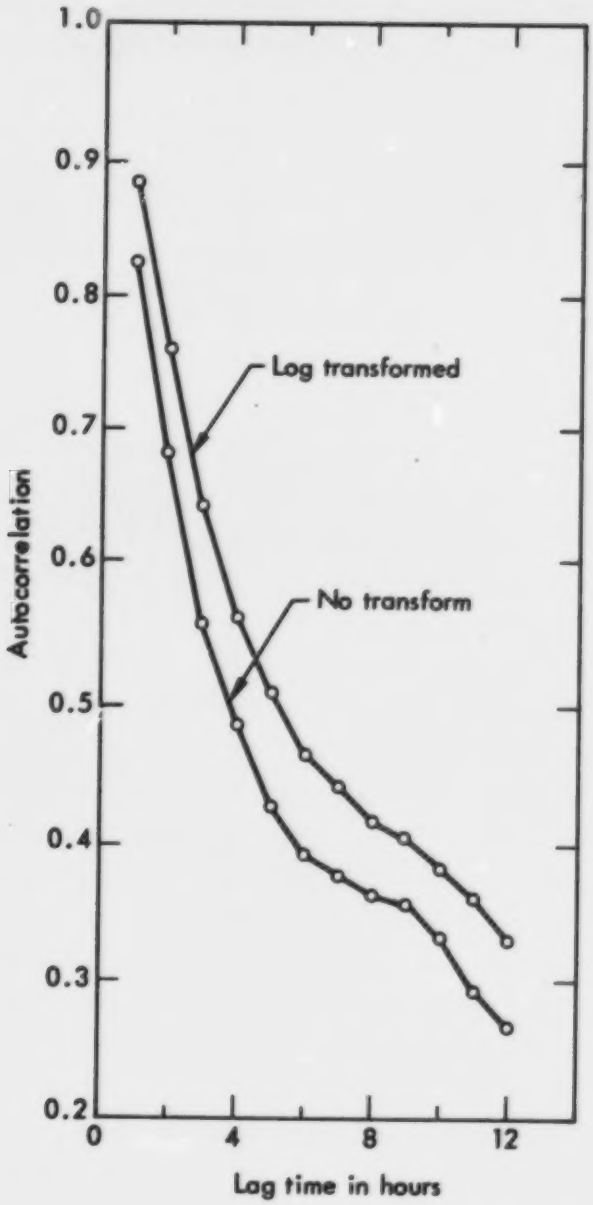

Fig. 1(a). Autocorrelation versus lag for CO concentrations in San Francisco, 1970 hourly averages.

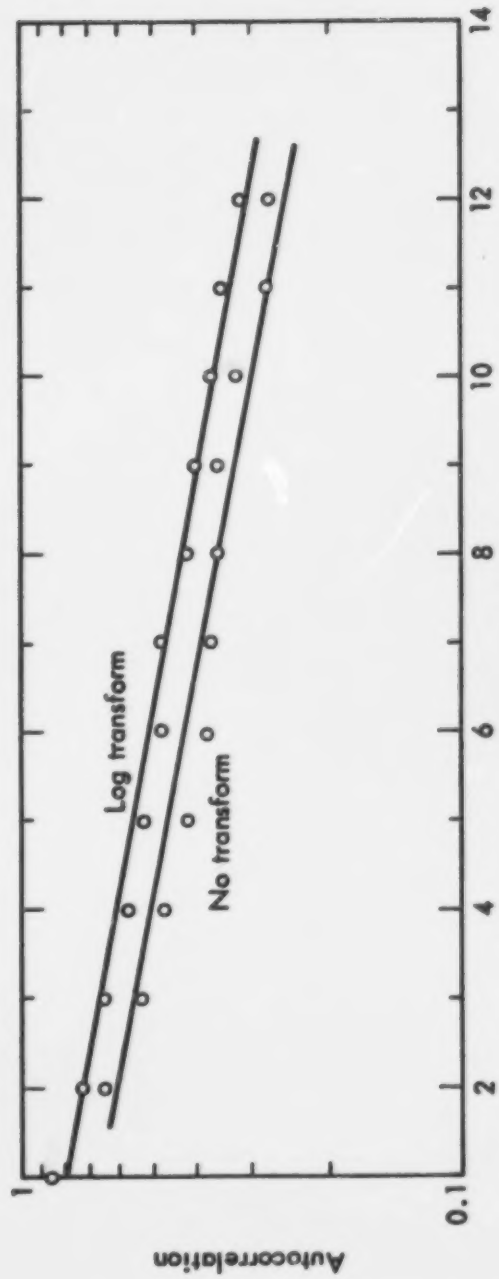

Fig. 1(b). Autocorrelation versus lag for CO concentrations in San Francisco, 1970 hourly averages.

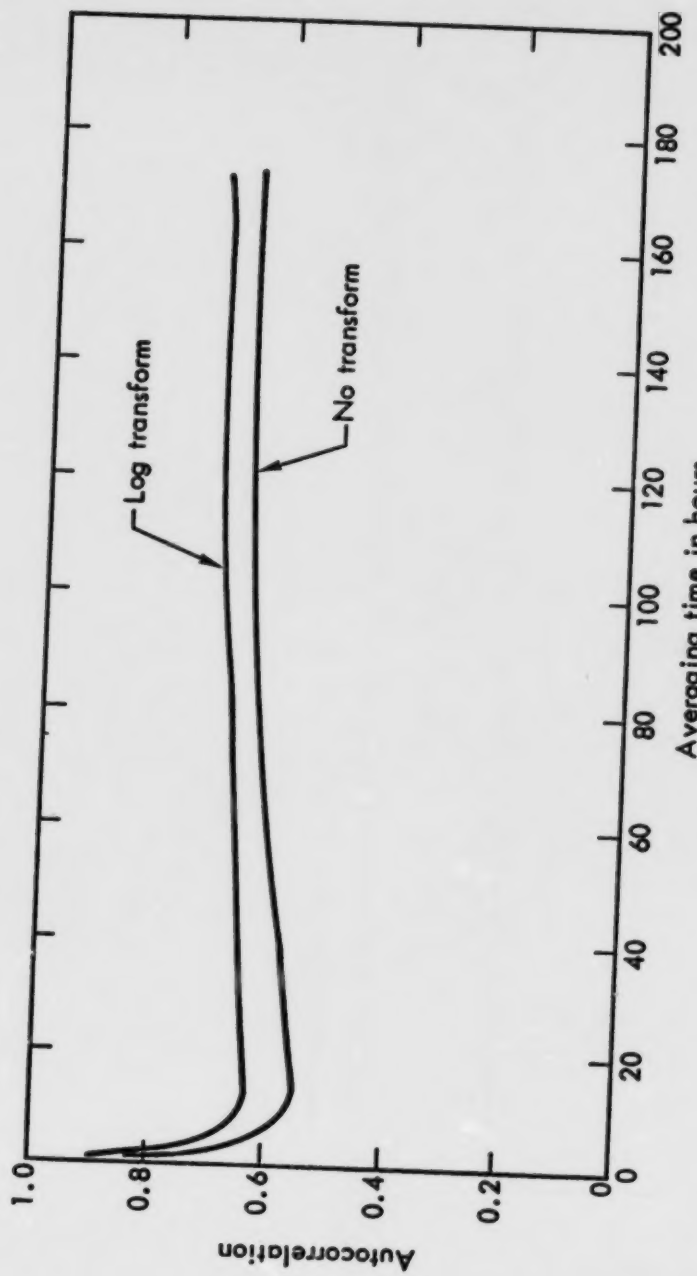

Fig. 1(c). Autocorrelation at lag 1 versus averaging time for CO concentrations in San Francisco, 1970 hourly averages.

Table 1. Autocorrelation and averaging time for CO concentrations in San Francisco, 1970 hourly averages.

Autocorrelation (lag 1)	Variance	Standard deviation	Mean	No. data points	Averaging time (days)
		Untransformed data			
0.834	3.943	1.986	3.488	20073	0.042
0.648	3.153	1.776	3.488	5018	0.167
0.576	2.285	1.512	3.488	1672	0.500
0.623	1.267	1.125	3.486	238	3.500
0.642	1.000	1.000	3.486	119	7.000
0.604	0.826	0.909	3.491	59	14.000
0.626	0.677	0.823	3.500	29	28.000
0.686	0.544	0.738	3.476	14	56.000
		Log-transformed data			
0.874	0.288	0.536	1.109	20073	0.042
0.705	0.243	0.493	1.109	5018	0.167
0.641	0.184	0.429	1.109	1672	0.500
0.675	0.116	0.340	1.109	238	3.500
0.706	0.095	0.308	1.109	119	7.000
0.664	0.081	0.285	1.110	59	14.000
0.664	0.068	0.261	1.112	29	28.000
0.726	0.059	0.243	1.104	14	56.000

Total number of hourly measurements = 20448

Number of missing measurements = 375

Table 2. Differences between autocorrelations of log-transformed and untransformed CO concentrations in San Francisco, 1970 hourly averages.

Autocorrelation	Averaging time (hr)	σ level of significance	Lag
0.834[a]	1	21.1	1
0.874[b]			
0.576[a]	12	4.23	1
0.641[b]			
0.626[a]	84	1.4	1
0.678[b]			
0.648[a]	168	1.3	1
0.712[b]			
0.681[a]	1	20.9	2
0.753[b]			
0.566[a]	1	18.9	3
0.650[b]			
0.487[a]	84	1.60	3
0.405[b]			

[a]No transformation of data.

[b]Natural log transformation of data.

Summary of Derivations

The latter derivation will serve as the basis for the reasoning presented later concerning the nature of the frequency distributions resulting from various types of sources and pollutants.

The other derivation supports the latter one, as is to be expected since the Gaussian Plume equation is a solution to the Fickian Diffusion equation. It is included for the sake of completeness and to demonstrate the consistency of the new derivation with existing theories.

The new derivation is based on reasoning which is clearer and more flexible than the Gaussian Plume derivation. These features will be used to great advantage in the explanation of various empirical results which could not be easily justified using the earlier derivation.

Other than these, no theoretical explanations for the lognormal distribution or any other have been suggested.

Conclusion

It is shown that considerable theoretical and empirical support exists for the lognormal distribution as the most appropriate for the characterization of pollutant concentrations for a wide range of averaging times. In later chapters other distributions are discussed, but these alternate results are demonstrated to be consistent with the material presented in this chapter.

CHAPTER III—FREQUENCY DISTRIBUTIONS OF RELATED VARIABLES

Pollutants can be viewed as tracers of atmospheric movements. Since we know that pollutant concentration frequency distributions are fit well by the lognormal, we suspect that some descriptors of the atmosphere are also lognormally distributed. Indeed, this is the case.

Fundamentally, atmospheric processes are structured differently than ordinary engineering-type processes. In particular, the change in a variable describing an atmospheric process is most often proportional to the level of that variable. This is written:

$$X_i = X_{i-1} \epsilon \, . \qquad \text{(III-1)}$$

This is a multiplicative process. Many descriptions of atmospheric motion can be described by a process of this form.

We are primarily interested in variables describing the transport of pollutant through the atmosphere. Such transport is described by the advective and eddy diffusive transfer rates. We are interested also in the removal of pollutants from the atmosphere, but few results are available except for particulates. Particle sizes, which partially govern the deposition of particulates, will be treated below.

Through this investigation of other meteorological variables we shall lend further credence to the conclusions presented in the preceeding chapter concerning the identity of the concentration distributions. In addition, the study of these variables yields greater insight into the nature of atmospheric motion and pollutant transport which leads to new approaches to identifying pollutant concentration frequency distributions.

Advection

The lack of knowledge concerning mesoscale atmospheric motions has hindered the formulation of a mathematical model describing such motion. For this reason it is difficult to demonstrate that windspeeds, measured continuously at a point over an interval of time, are lognormally distributed from theoretical considerations. However, extensive empirical analysis has been performed, indicating that the lognormal is a reasonable assumption.

It has been shown by Gifford and Hanna (11), (12) that pollutant concentrations are proportional to windspeeds, which is also implied from the fact that both are lognormally distributed (see Fig. 2). The correlation coefficient for nonreactive primary pollutants like CO is extremely high (~0.90), slightly less for the more reactive pollutants like NO_2 (~0.85), and least for the secondary pollutants like oxidant (0.66). These investigators evaluated the constants of proportionality for many cities for several pollutants. The results are summarized in Refs. (11) and (12).

In comparative studies, it has been found that this simple model

$$\psi = KQ/U \qquad \text{(III-2)}$$

where

Q = total emissions,

U = windspeed,

ψ = pollutant concentration and

K = empirical constant

performs as well as many more complicated models for primary pollutants. This is a testimonial to the fact stated above, i.e., pollutants are tracers of atmospheric motion. Although this may seem obvious, one should realize that other processes than just advection influence concentration, including the time history of the source, the deposition, and the distance the pollutant has been transported.

Diffusion

The windspeed is a measure of advection, while diffusion is described by eddy diffusivities. In the Fickian diffusion equation,

$$\frac{d\psi_a}{dt} + u\frac{\partial \psi_a}{\partial x} + v\frac{\partial \psi_a}{\partial y} + w\frac{\partial \psi_a}{\partial z} = \frac{\partial}{\partial x}\left(K_x \frac{\partial \psi_a}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y \frac{\partial \psi_a}{\partial y}\right)$$

$$+ \frac{\partial}{\partial y}\left(K_z \frac{\partial \psi_a}{\partial z}\right) + \frac{S_a(x,y,z,t)}{V} + \frac{P}{V}(\psi_a, \psi_b \cdots \psi_n, t), \qquad \text{(III-3)}$$

K_x, K_y, K_z are constants which are the eddy diffusivities describing diffusive flux. We may suspect that these are also lognormally distributed, and the process of demonstrating this is now presented.

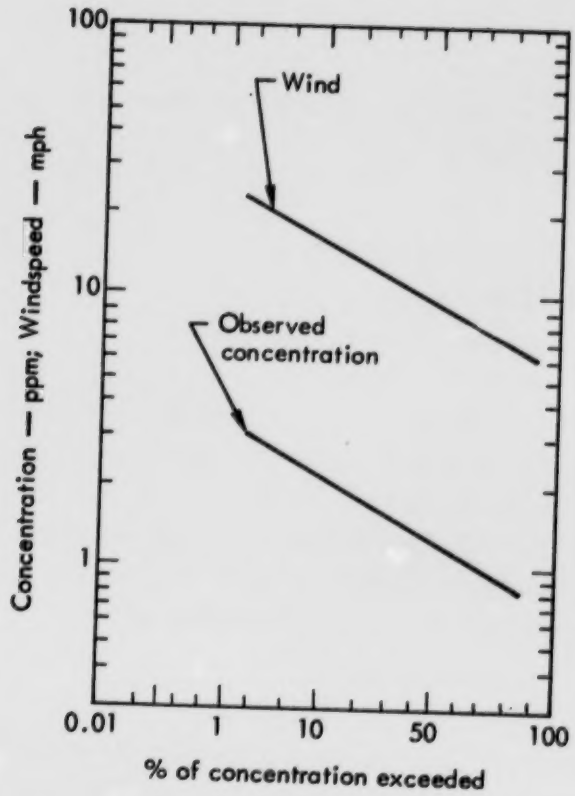

Fig. 2. Concentration and windspeed frequency distributions for CO and windspeed for San Francisco, 1970 hourly averages.

The local viscous dissipation is turbulent flow, ϵ, given by

$$\epsilon = \frac{\nu}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)^2 \qquad \text{(III-4)}$$

where ν is viscosity, u is velocity, and i and j refer to direction.

In his original similarity hypothesis, Kolmogoroff formed length and time scales of turbulent motion without taking the variability of ϵ into consideration. In 1962, Kolmogoroff refined this hypothesis to take into account random fluctuations in ϵ (13). In turbulent flow, energy is transferred from one stage of motion to the next; this transfer is what ϵ represents. The amount of energy transferred at each stage has been argued to be a function of the relative magnitude of the state (14). This describes a multiplicative process. If the transfer stages are similar and are independent, then by the law of proportional effect, $\ell n\,(\epsilon)$ is distributed normally.

Now, recalling the reproductive properties of the distribution we can argue the lognormality of several other variables for which the relations to ϵ are well known. The dissipation of temperature variance by thermal conduction is given by

$$X = 2\alpha(\text{grad } T)^2 \qquad \text{(III-5)}$$

where α is thermal diffusivity. If both X and ϵ are lognormal [X is argued lognormal by Gurvich (15)] then ΔT and ΔU, the temperature and velocity differences between two points in space separated by a distance, are lognormal on either side of the origin, or more concisely, $(\Delta T)^2$ and $(\Delta U)^2$ are lognormal due to the relationships:

$$(\Delta T)^2 \approx X_1 \epsilon_r^{-1/3} r^{2/3}$$

$$(\Delta U)^2 \approx \epsilon_r^{2/3} r^{2/3} \qquad \text{(III-6)}$$

from Kolmogoroff and the reproductive properties of the distribution.

Measurements (15) for r = 2 cm at 4 m above ground yielded the distribution function plotted on lognormal probability paper in Fig. 3. Clearly, the lognormal is a good approximation where a straight line indicates perfect lognormality.

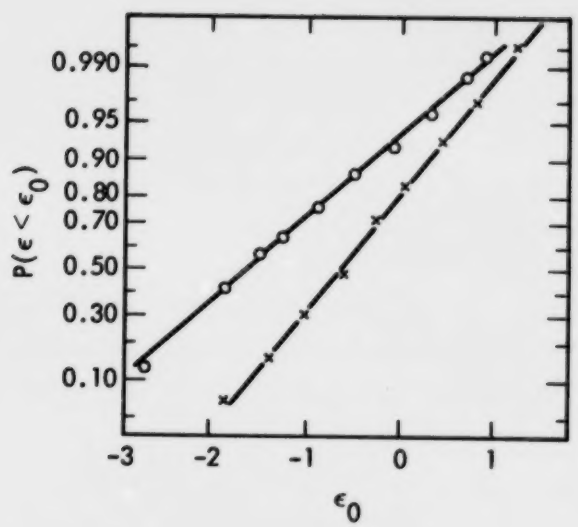

Fig. 3. Probability distribution of the squared temperature difference compared with lognormality $P(\epsilon < \epsilon_0)$. $\epsilon = (\Delta T)^2 / [(\Delta T)^2](15)$. Separation = 2 cm, 10^4 samples per plot.

An approximation to the horizontal eddy diffusion coefficient based upon similarity theory is (16)

$$K_h = \epsilon^{1/3} \sigma^{4/3} \qquad \text{(III-7)}$$

where σ is the root-mean-square dispersion of the particles in a pollutant puff.

Since a lognormal variable raised to any exponent is also lognormal and since $\sigma^{4/3}$ is invariant, it can be inferred that K_h is lognormally distributed. Note that σ has been approximated at $0.7\ \Delta S$ where ΔS is an intergrid-square distance in a compartmentalized model.

Further, another approximation which is used in atmospheric modeling is

$$\epsilon = \frac{300}{Z}\left(\frac{u}{5}\right)^3 \qquad \text{(III-8)}$$

which was developed without reference to statistical considerations. This relationship indicates the lognormality of windspeed based upon the lognormality of ϵ, or vice versa (17). Also, the vertical eddy diffusivity has been approximated as

$$K_z = 400\ u_1 \qquad \text{(III-9)}$$

where u_1 is the horizontal wind velocity at 1 meter, determined by use of a power law vertical profile from the mean layer wind. This illustrates the likelihood that K_z, a quantity which has not yet been accurately measured in the atmosphere, is lognormally distributed.

It is worthwhile to notice that the results obtained from models using these approximations have been encouraging (8).

Particle Sizes

It has been clearly established (18) that particle size distributions in atmospheric aerosols are approximately lognormal. There is some disagreement between scientists as to the exact size range covered, or whether there are two or three overlapping lognormal distributions (19). However, the lognormal approximation is widely accepted.

Particles are the result of a multistage grinding process where the size of a particle at any stage is a function of its size at the immediately previous stage. This can be represented by the multiplicative process which results in the lognormal distribution examined in Chapter II.

The deposition rate of the particles is a function of particle size, primarily, and is therefore approximately lognormally distributed by virtue of the reproductive properties of the distribution. Therefore, the negative portion of the source term in Eq. (II-3) due to deposition is approximately lognormally distributed.

Conclusion

Pollutant concentrations are a function of emissions, chemical change, deposition and transport. Several of the variables describing transport and deposition are discussed here and are found to be consistent with the lognormal assumption.

Hence it is clear that pollutant concentrations are followers of the overall character of motion in the atmosphere. This result has not yet received adequate attention. Full realization of its significance indicates that air quality data, which is essentially simple to collect and useful for a pragmatic purpose, may have applications in the understanding of atmospheric processes.

CHAPTER IV—FREQUENCY DISTRIBUTIONS FOR VARIOUS POLLUTANTS AND SOURCE TYPES

In this chapter we shall examine several aspects in which pollutants differ, in light of the model presented in Chapter II. It will be shown that certain seemingly conflicting results can be explained in this context.

Reactive versus Inert Pollutants

When we examine frequency distributions of air pollutant concentrations, we notice that the parameters vary from day to day and from pollutant to pollutant. This variation is a result of the nature of the meteorology of which the pollutant is a tracer, the sources of the pollutant, and its reactivity.

Figures 4(a) and 4(b) show daily profiles of several pollutants. The profiles with the steepest slopes and highest peaks yield the highest standard geometric deviation (SGD). We can see that pollutants with similar reactivity trace the meteorological conditions in the same way, provided that they are from the same type of source. Carbon monoxide and hydrocarbon are an example of this. Shuck et al. (20) report a correlation coefficient of 0.99.

These principles are simple to see; anything which causes large fluctuations in the daily profile will cause a large SGD in the frequency distribution of that pollutant. The significant causes are unstable meteorological conditions and volatile pollutants. This volatility is caused by either the basic chemical structure of the pollutants, as is the case with oxidants and oxides of nitrogen, or by the temperature of the pollutant, as is the case with thermal plumes of SO_2.

If we return to the Fickian Diffusion equation:

$$\frac{d\psi_a}{dt} + u\frac{\partial\psi_a}{\partial x} + v\frac{\partial\psi_a}{\partial y} + w\frac{\partial\psi_a}{\partial z} = \frac{\partial}{\partial x}\left(K_x\frac{\partial\psi_a}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y\frac{\partial\psi_a}{\partial y}\right)$$

$$+ \frac{\partial}{\partial y}\left(K_z\frac{\partial\psi_a}{\partial z}\right) + \frac{S_a(x,y,z,t)}{V} + \frac{P}{V}(\psi_a, \psi_b \ldots \psi_n, t) \qquad \text{(IV-1)}$$

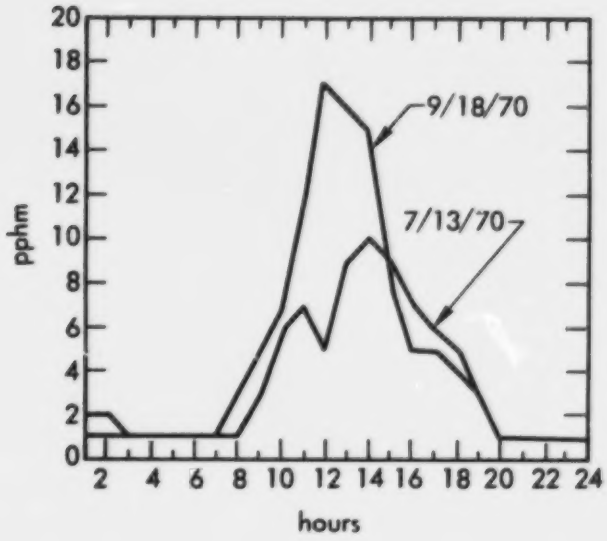

Fig. 4(a). Oxidant concentration versus time in Los Angeles, hourly averages.

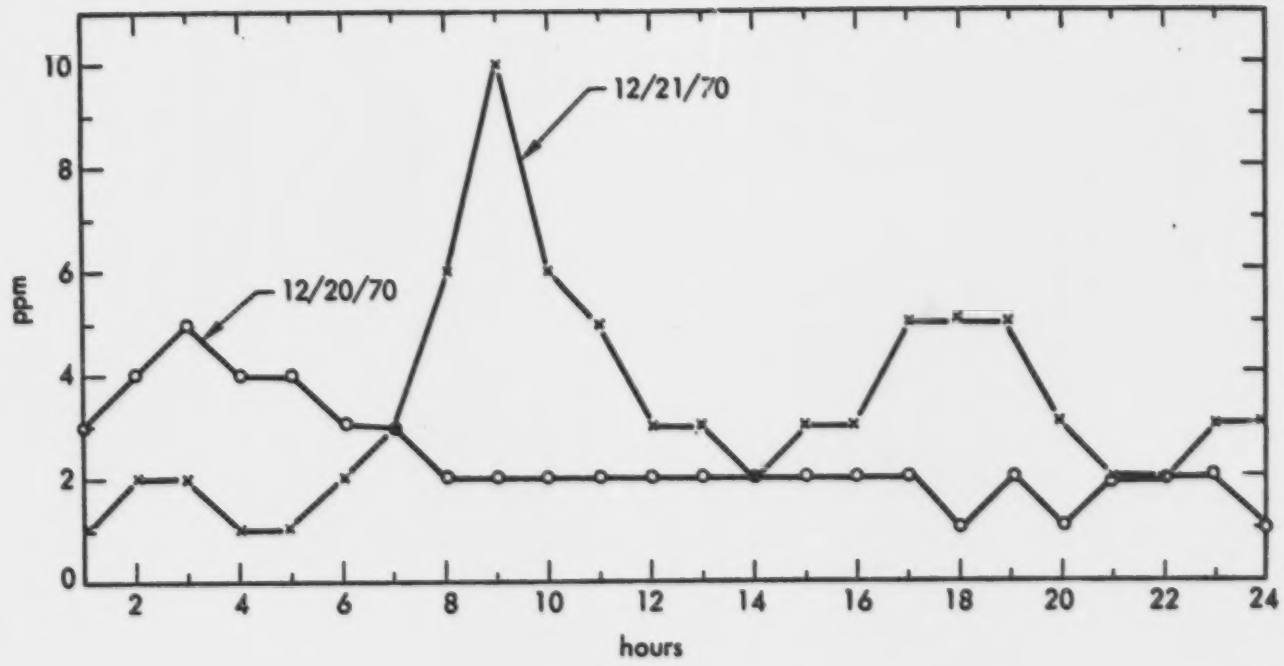

Fig. 4(b). CO concentration versus time in San Francisco, hourly averages.

we see that the chemical change term is larger and more variable for reactive pollutants. This affects the argument presented in Chapter II through its direct effect on the daily profile predicted from the equation. The increased volatility causes larger fluctuations in concentration which are seen in the box model diffusion [Eq. (II-18)] both through the magnitudes of $\psi_k(m, b)$ and $\psi_k(j,t)$, and in the magnitude of the difference $E_k(j, t)$. This absolute value of this term is larger, but it still changes sign over the diurnal cycle which results in lognormality (Fig. 5).

This explains the results of Benarie (3) and Knox and Lange (2) who noticed changes in SGD between pollutants for the same meteorology. The model described in Chapter II provides a theoretical framework through which these results can be understood.

Although Larsen's analysis indicated that oxidant concentrations are lognormally distributed for all cities for all averaging times as in Fig. 6(a) often oxidant concentrations depart from lognormality in a consistent way, as is depicted in Fig. 6(b). At the present time, no definitive explanation has been set forth to explain this anomaly. A brief investigation has suggested two nonmutually exclusive, possible explanations.

The first is based upon the theory that the photochemical reactions producing oxidant in the atmosphere are self-limiting. This is given some support by the fact that oxidant concentrations have never been recorded at 1 ppm or higher in the atmosphere, even when such concentrations might be expected because of source and meteorological conditions.

The second explanation is that the averaging time of 5 minutes is so long in comparison to the reaction rates that it averages out short-duration high concentrations which might otherwise cause the cumulative distribution curve, as in Fig. 6(b) to straighten out. This is supported by the fact that the "angle" between the two adjacent straight lines in Fig. 6(b) becomes sharper as averaging time increases, and there is no reason why the reverse should not hold for averaging shorter than 5 minutes.

The shape of the curve in Fig. 6(b) can be compared to the shape of the curves for the gamma, Weibull, and Pearson-IV distributions plotted on log probability paper. It appears that the Weibull distribution

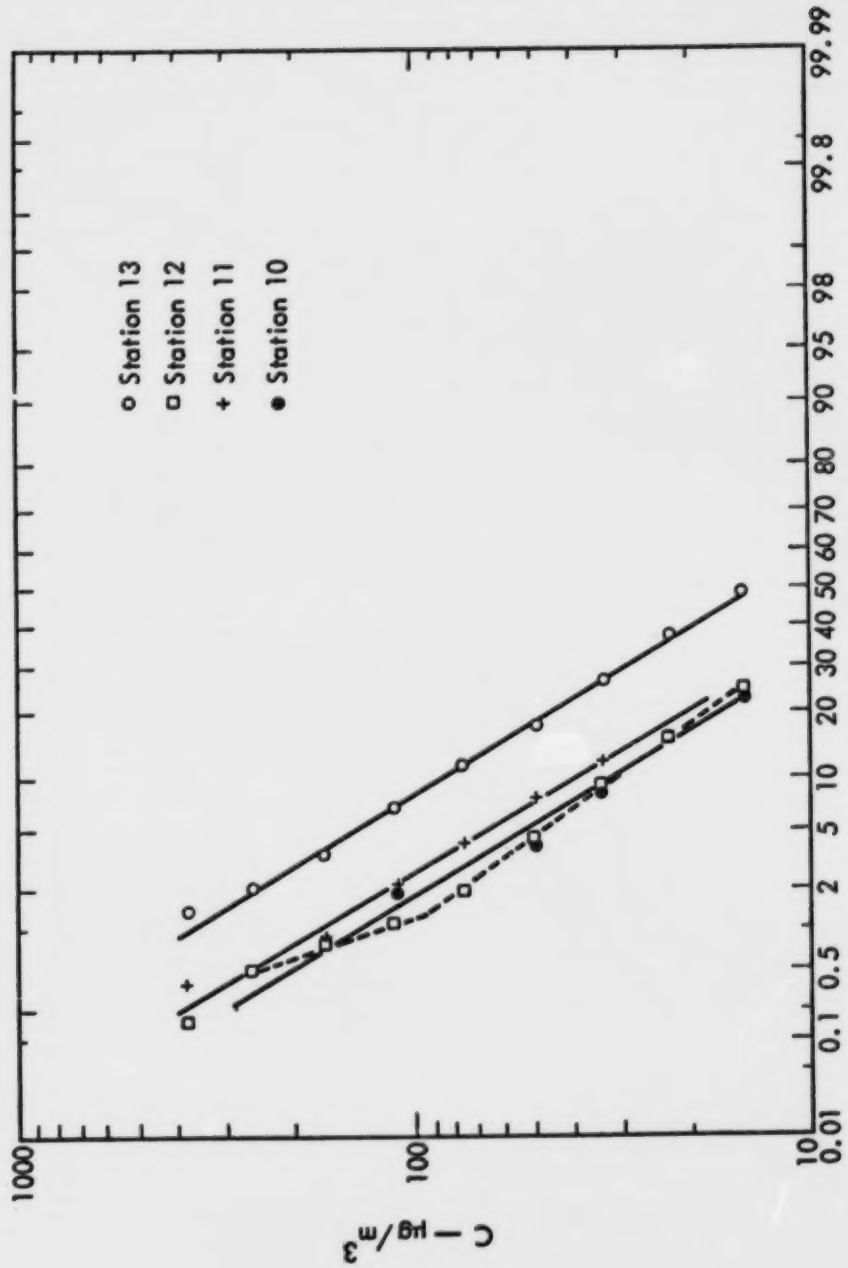

Fig. 5(a). SO_2 concentrations, direction 289° – 308°, Lacq, France, 1968 – 1969 (3).

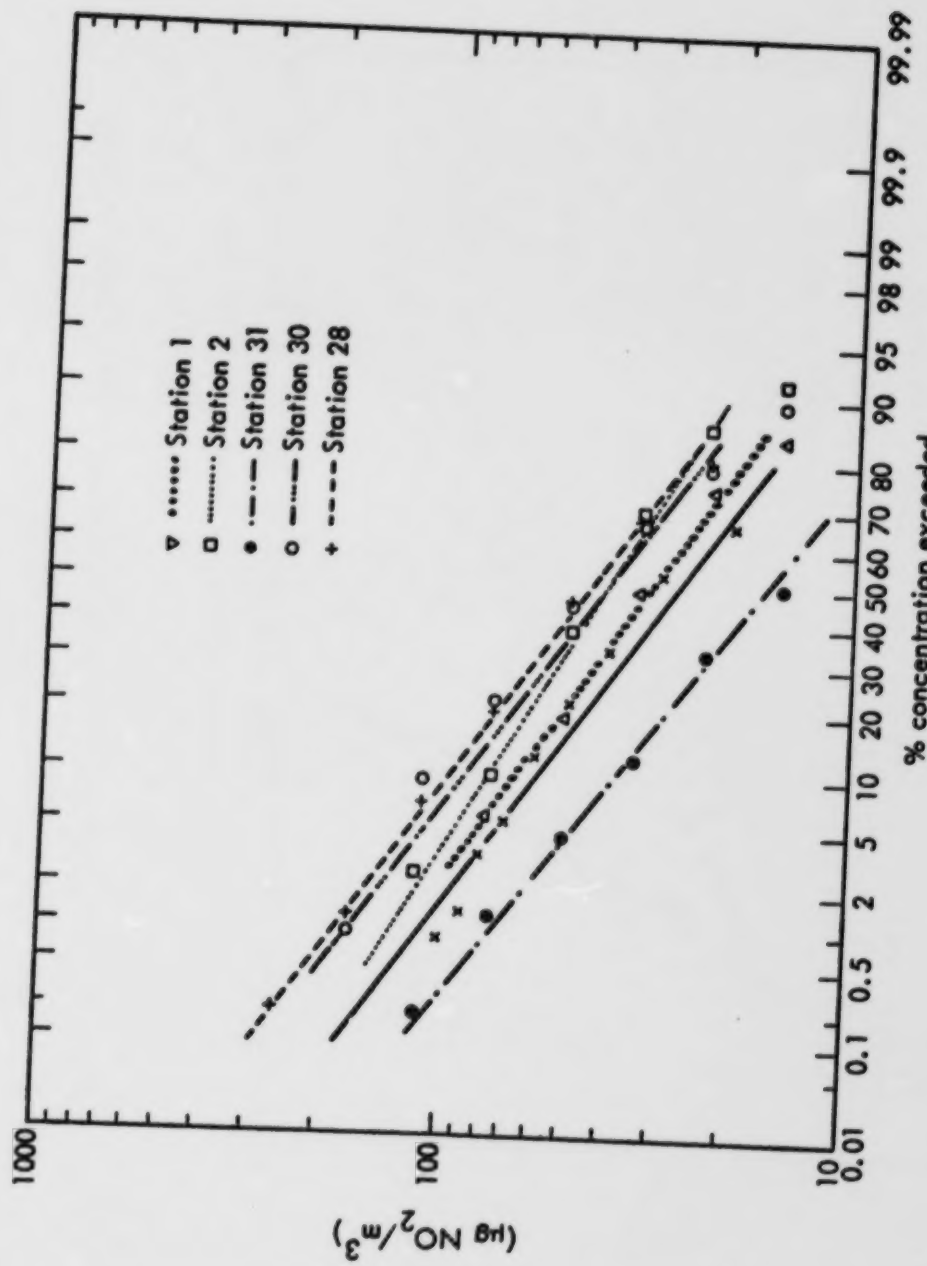

Fig. 5(b). NO_2 concentrations, direction 108° - 121°, Lacq, France, 1968-1969 (3).

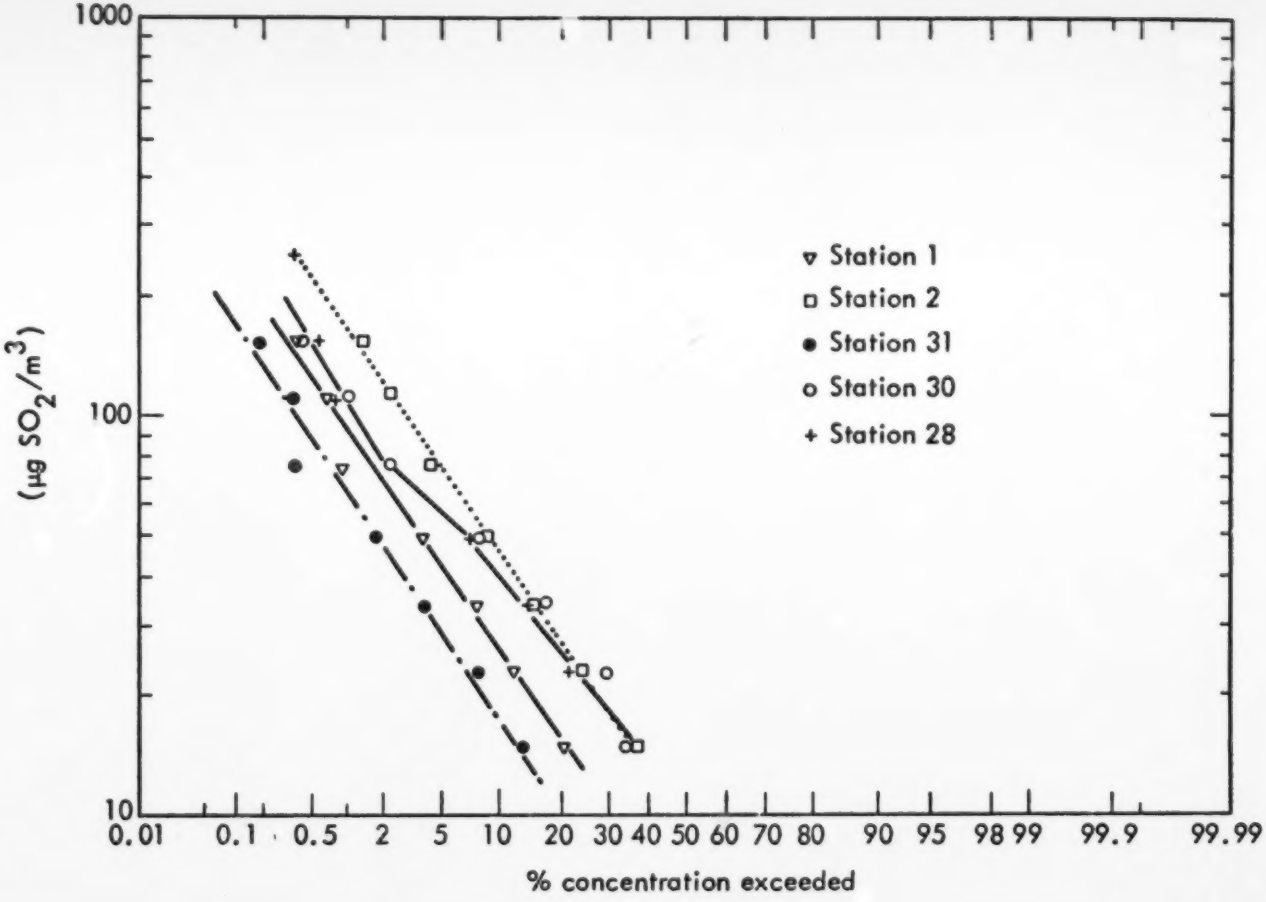

Fig. 5(c). SO_2 concentrations, direction 108° - 121°, Lacq, France, 1968 - 1969 (3).

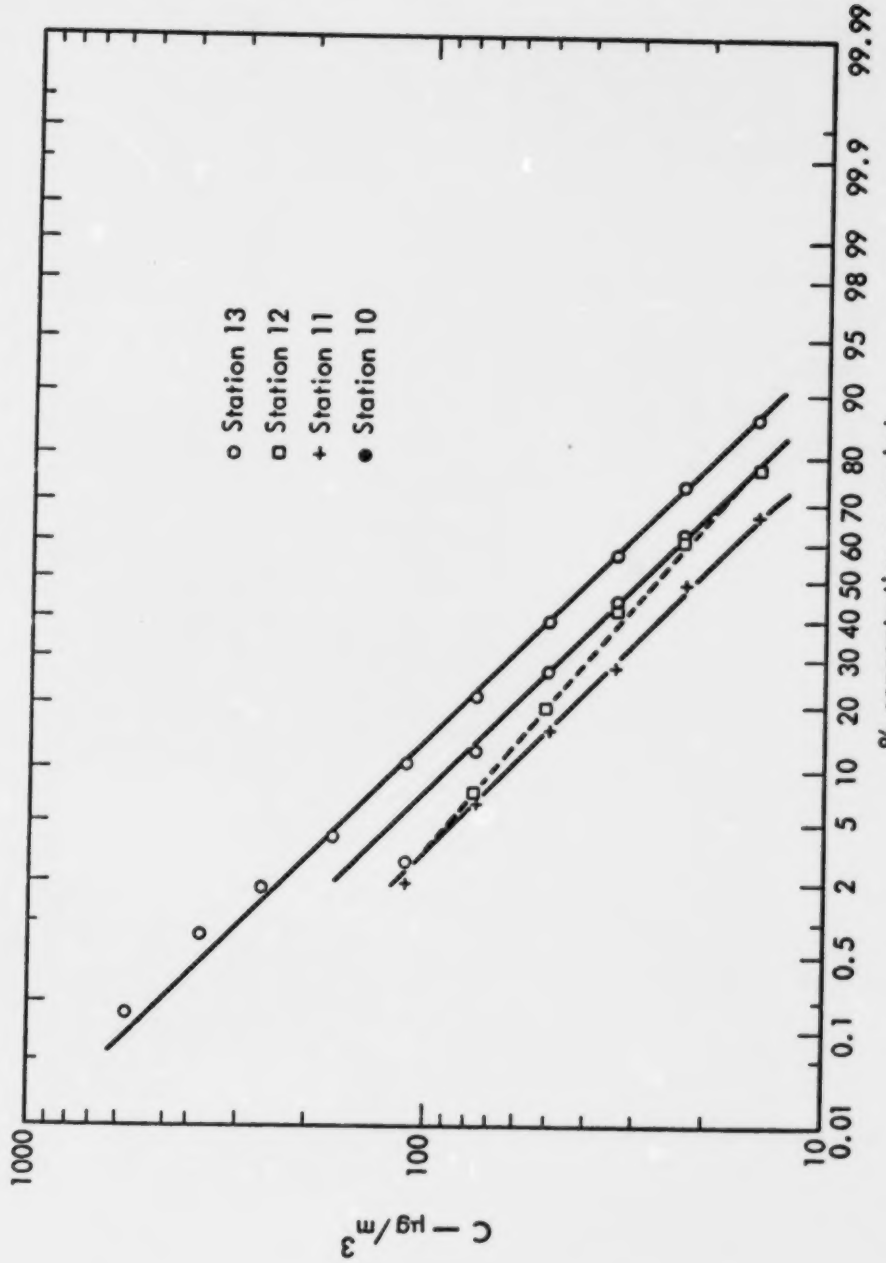

Fig. 5(d). NO_2 concentrations, direction 289° - 308°, Lacq, France, 1968 - 1969 (3).

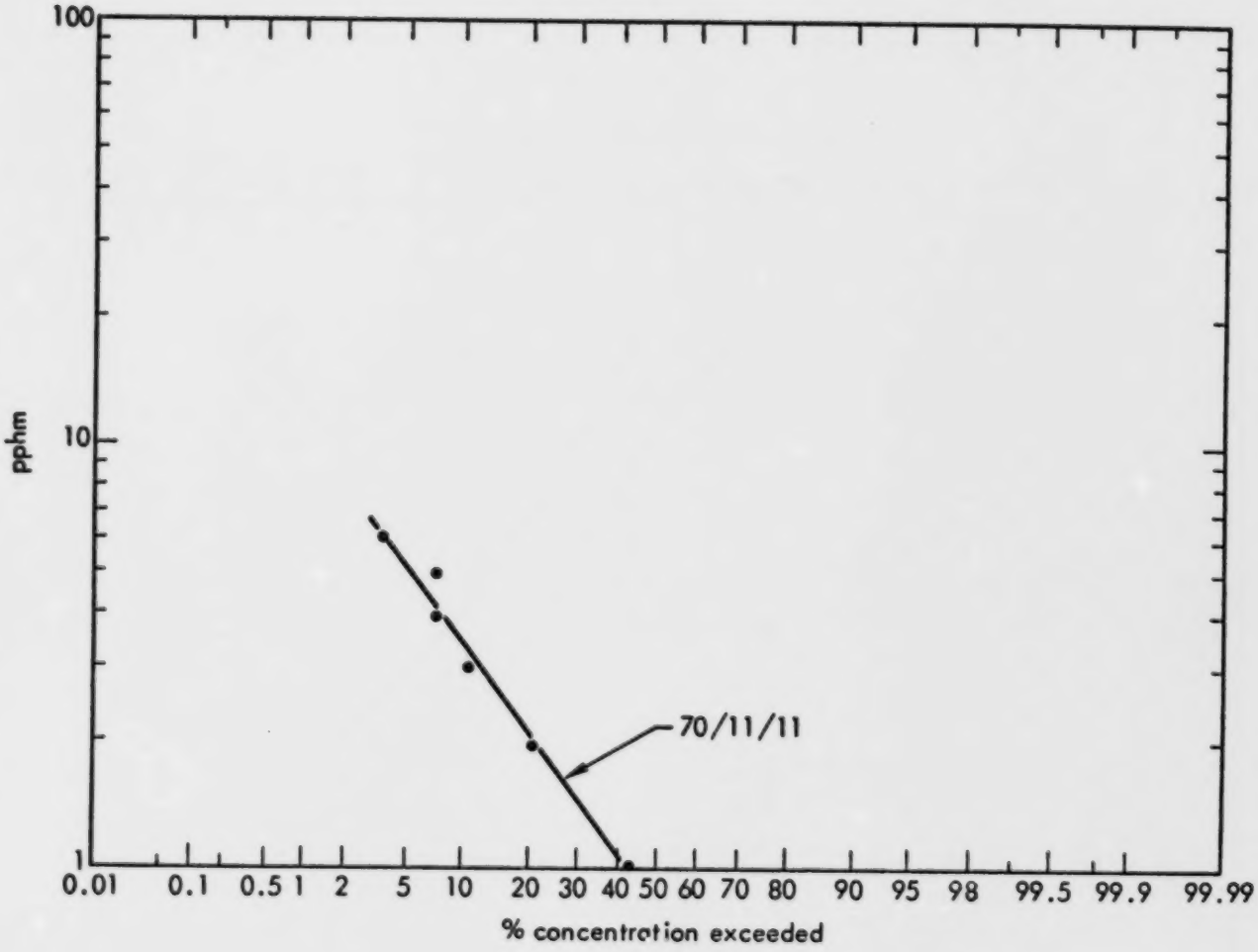

Fig. 6(a). Log probability plot of oxidant concentrations in Los Angeles, 11/11/70, hourly averages.

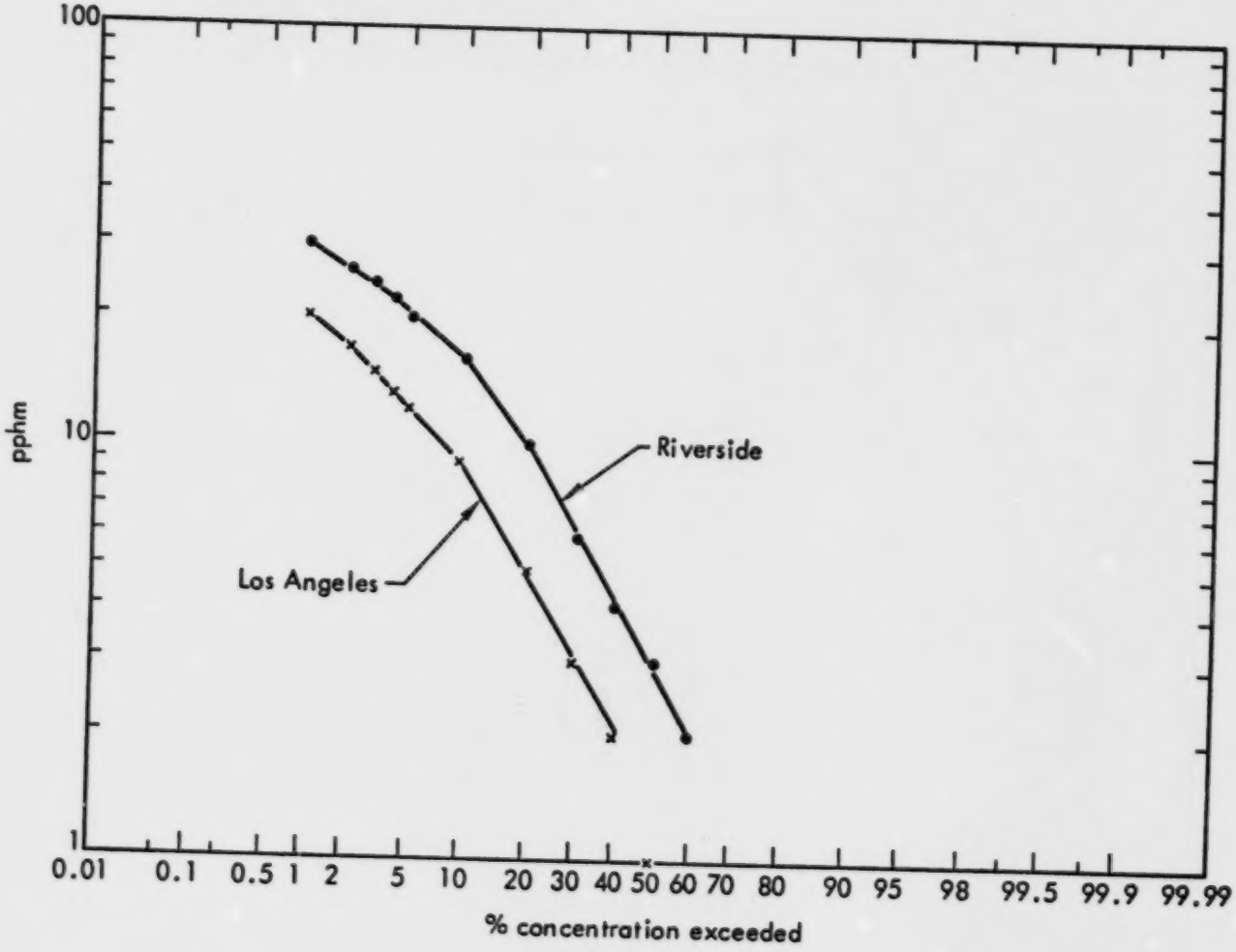

Fig. 6(b). Oxidant concentrations in Riverside California and Los Angeles California, 1967 hourly averages.

would provide a better fit, although there is no apparent theoretical reason for this to be the case. The beta and gamma distributions do not appear to fit the tail of the oxidant concentration distribution at all well. The nature of these distributions is discussed more extensively in Chapter V.

Point versus Area Sources

The question of the difference in concentration distributions between point and area sources is of great interest from a practical standpoint. One must be able to evaluate the contribution of a large point source toward air pollution to determine, for example, the feasibility of a particular location for a polluting industry. It is, however, a difficult question for which the literature contains conflicting answers.

Gifford (5) has proved theoretically and has presented a limited amount of data to the effect that logs of concentrations of a pollutant from a point source are proportional to a chi-squared distribution for suitably normalized data based on the Gaussian Plume equation. Benarie (3) has proven that such concentrations are lognormally distributed on the basis that they are merely tracers of a lognormally distributed windfield. Knox and Lange (2) analyzed a 5-year release of Argon-41 from the Chalk River reactor in Ontario, Canada, (Argon-41 has a zero background concentration) and determined that the lognormal distributions fit poorly, although they did not propose an alternate distribution. An analysis of their data indicates that the chi-squared distribution did not fit well either.

In Chapter II a modified version of the derivation of the frequency distribution from a point source is presented. The result of this derivation, which considers the variability in the windfield, is a distribution composed of a sum of chi-squared and lognormal components determined by the magnitude and direction of the windfield, the stability conditions, and the distance from the source.

In the cases presented in the literature, these factors are not controlled. There is not yet enough data to determine the distribution identity as a function of the relevant variables. The point to be made here is that the presented equations do indeed predict such differences as

are reported, and that future research may yield a quantitative treatment of these differences.

A final point is that the diffusion-equation-based model is consistent with this result. The form of the predicted frequency distribution is dependent on the relative source strength, and is sensitive to stability conditions and windspeeds through the advective and eddy diffusive transfer fluxes. However, the exact form is more difficult to predict from this model, which is more appropriate for area sources.

Summary

This chapter discusses the differences in pollutant concentrations resulting from point and area sources. It also discusses fundamental differences in the distributions measured for reactive and inert pollutants.

These differences are explained within the framework of the models discussed in Chapter II. The fact that the observed distributions appear to be consistent with model predictions lends further support to the validity of the concepts presented here.

Lognormal Distribution

The natural logarithm of a lognormally distributed random variable is normally distributed. This relationship implies that the lognormal is the multiplicative analog to the normal distribution. In particular, where the process

$$x_i = x_{i-1} + \epsilon \tag{V-1}$$

generates a normally distributed random variable, the process

$$x_i = (x_{i-1})\,\epsilon \tag{V-2}$$

generates a lognormally distributed random variable; ϵ is an arbitrarily distributed random shock.

Indeed, many physical processes are best described by Eq. (V-2) and hence their result is lognormally distributed. The lognormal is more than a variation of the normal distribution, in fact it is one of the most fundamental distributions of mathematical statistics. Increasingly it is being found that the outputs of physical processes are lognormally distributed. It is a distribution that physicists, meteorologists and engineers all encounter.

The lognormal distribution is given by:

$$\Lambda(Y) = N(\log x) \qquad x > 0 \tag{V-3}$$

and

$$d\Lambda(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(\log x - u)^2\right] dx \qquad x > 0.$$

The jth moment about the origin is given by

$$m_j = \int_{-\infty}^{\infty} x^j \, d\Lambda(x) \tag{V-5}$$

$$= \int_{-\infty}^{\infty} e^{jy} \, dN(y) \tag{V-6}$$

$$= e^{ju + 1/2 j^2 \sigma^2} . \qquad \text{(V-7)}$$

Therefore, the mean and variance are given by

$$\alpha = e^{u+1/2\sigma^2} \qquad \text{(V-8)}$$

$$\beta^2 = e^{2u+\sigma^2}\left(e^{\sigma^2} - 1\right) = \alpha^2\eta^2 \qquad \text{(V-9)}$$

where $\eta^2 = e^{\sigma^2} - 1$. The third moment is:

$$m_3 = \alpha^2(\eta^6 + 3\eta^4) \qquad \text{(V-10)}$$

and the fourth moment is

$$m_4 = \alpha^4(\eta^{12} + 6\eta^{10} + 15\eta^8 + 16\eta^6 + 3\eta^4) \qquad \text{(V-11)}$$

which results in nonzero coefficients of skewness S_1 and kurtosis S_2,

$$S_1 = \frac{m_3}{\beta^3} = \eta^3 + 3\eta \qquad \text{(V-12)}$$

$$S_2 = \frac{m_4}{\beta^4} - 3 = \eta^8 + 6\eta^6 + 15\eta^4 + 16\eta^2 \qquad \text{(V-13)}$$

Skewness and kurtosis are both positive and both increase as the variance increases.

The mode of the distribution is given by $e^{u-\sigma^2}$, the median by e^u, and the mean by $e^{u+\sigma^2/2}$, hence the curve appears as in Fig. 7. Figures 8 and 9 illustrate the effect of varying the parameters.

Most important to the present studies are the reproductive properties of the distribution. The necessary theorems will be stated with outlines of the proofs as required.

Theorem 1. If x_1 and x_2 are indpendent Λ variates, then the product x_1x_2 is also a Λ variate.

Proof: This is proved by taking logs to convert the distributions to normal distributions, convolve the resulting variables and convert the result back using an antilog transform.

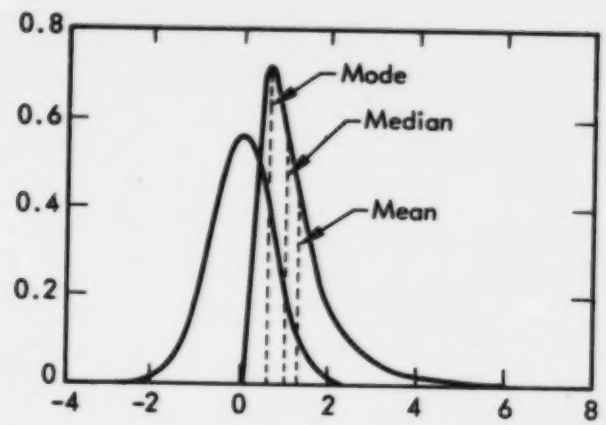

Fig. 7. Frequency curves of the normal and lognormal distributions.

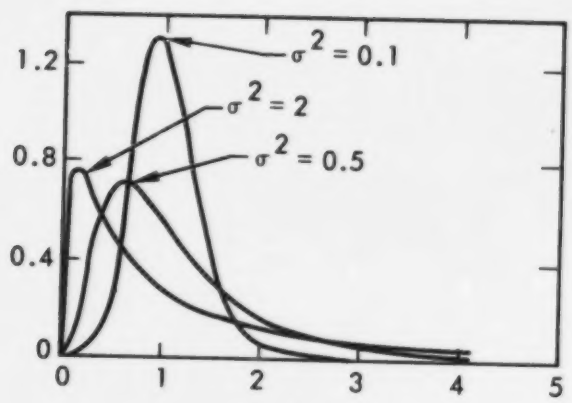

Fig. 8. Frequency curves of the lognormal distribution for three values of σ^2.

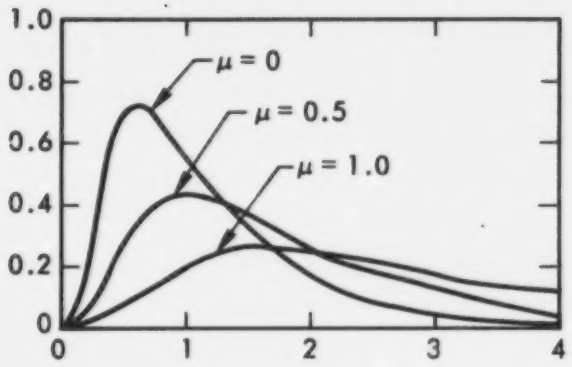

Fig. 9. Frequency curves of the lognormal distribution for three values of μ.

Theorem 2. If $\{x_j\}$ is a sequence of independent positive variates having the same probability distribution and such that:

$$E\{\log x_j\} = u \tag{V-14}$$

$$V^2\{\log x_j\} = \sigma^2 \tag{V-15}$$

and both exist, then the product $\prod_{j=1}^{n} x_j$ is asymptotically distributed as $\Lambda(nu, n\sigma^2)$.

Proof: By analogy to the additive normal Central Limit Theorem.

For limited numbers of variates in the sum it has been demonstrated that the sum variable is lognormally distributed for certain ranges of the coefficient of variation (21). Figure 10(a) gives these conditions.

Goodness-of-fit tests have been derived for the lognormal. The chi-squared test is appropriate of course; the Kolmogoroff-Smirnov test is a nonparametric technique for determining a confidence band around an empirical distribution function. Another useful test is to plot the data on lognormal probability paper on which truly lognormal data will plot as a straight line. Figures 10(b) and 10(c) illustrate several such plots.

Up to this point we have discussed primarily the lognormal distribution which appears to have considerable empirical and theoretical support. There are, however, other distributions which fit the same data quite well and are therefore deserving of mention. It is interesting to note, however, that no non-empirical support for these distributions has yet been published.

Lynn (22) used data from Philadelphia, Pennsylvania to estimate the parameters of several distributions by the method of moments. The distributions are the normal, two-parameter lognormal, three-parameter lognormal, gamma, and Pearson-IV parameter. The goodness-of-fit statistics are summarized in Table 3.

Not considered here is the Weibull distribution which has considerable support from several sources, Milokaj (23) and Barlow (24).

In Table 3, notice that the normal distribution was clearly the worst. The two-parameter lognormal was the best by a small margin

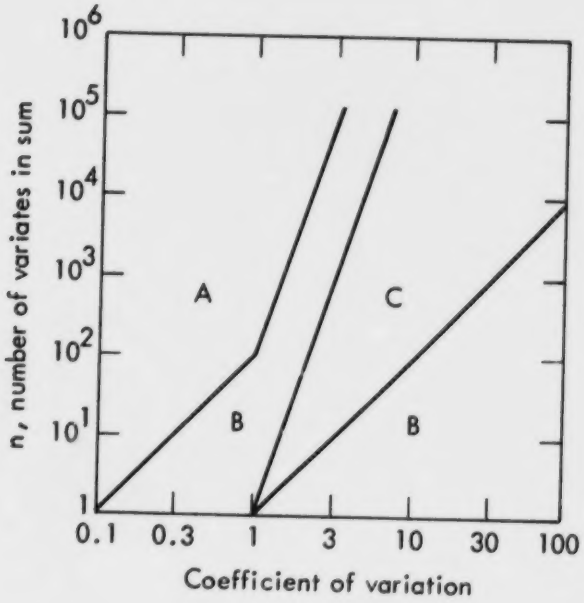

Fig. 10(a). Regions of convergence where the sum of n lognormal variates is approximately lognormal. (A) Convergence for both normal and lognormal approximations, (B) convergence for the lognormal approximation, (C) convergence is uncertain (21).

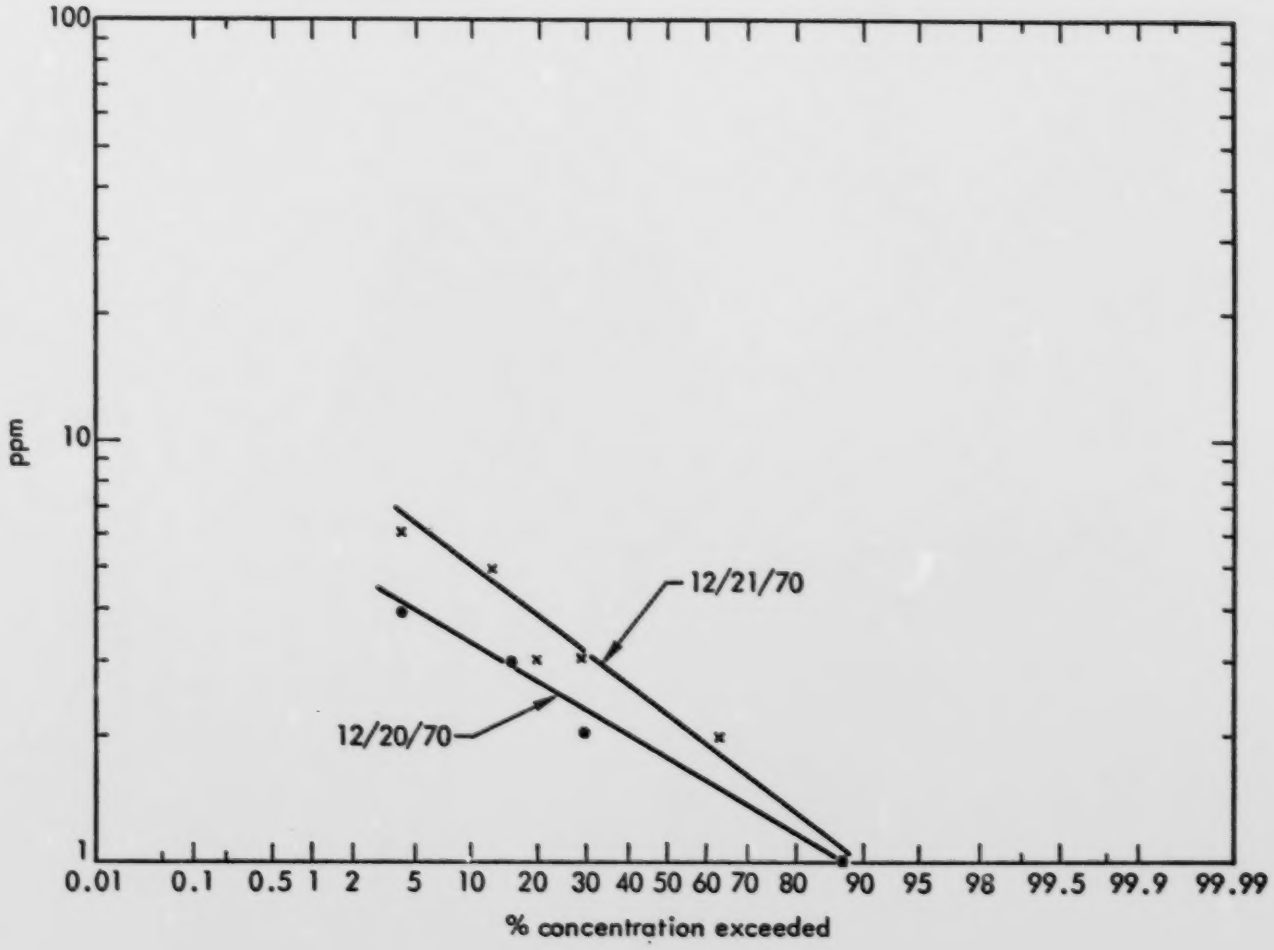

Fig. 10(b). CO concentration in San Francisco, hourly averages.

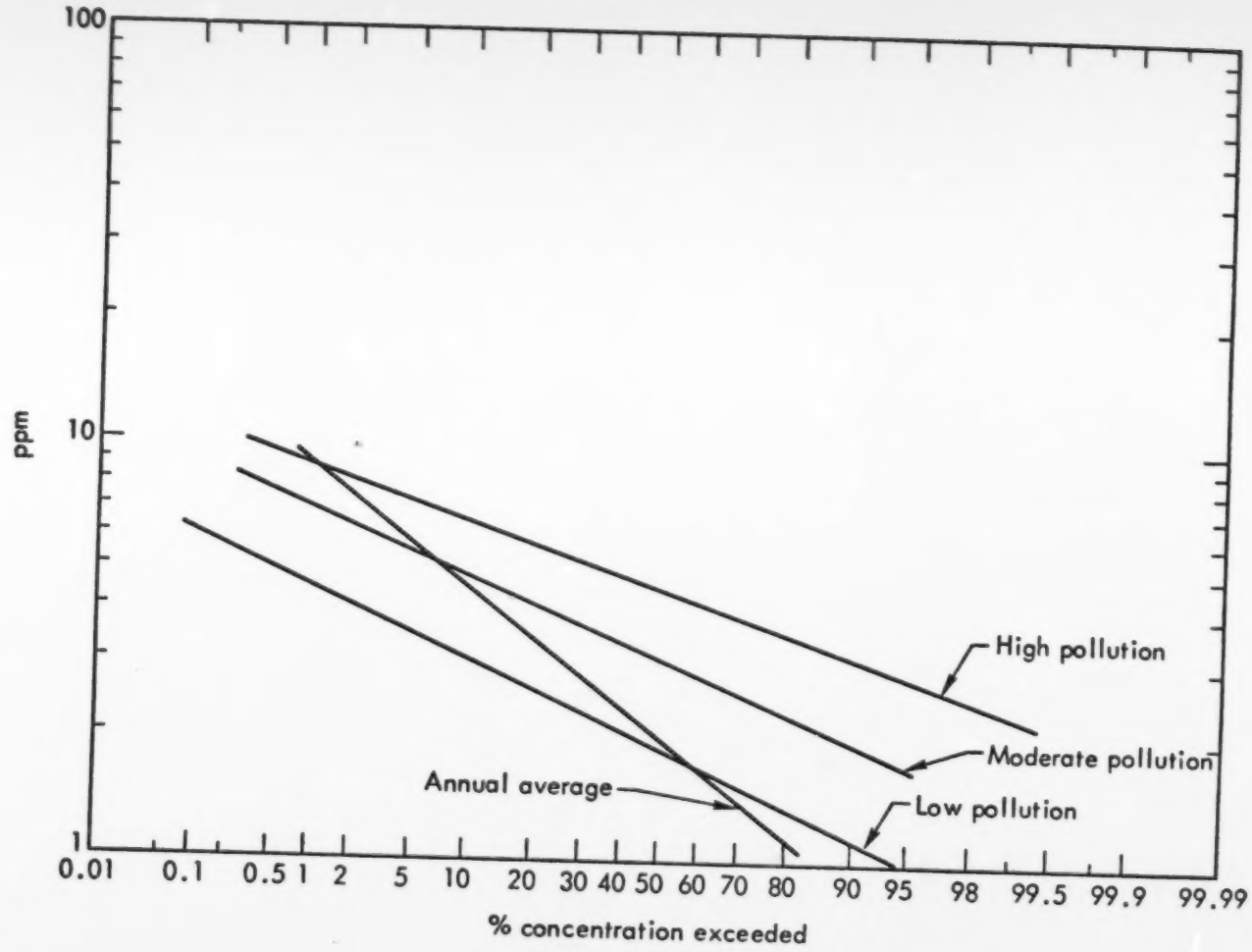

Fig. 10(c). CO concentrations for various categories of pollution days in San Francisco, 1970 hourly averages.

Table 3. Summary of total absolute deviations (20 $\mu g/m^3$ classes).[22]

		Normal dist.	Lognormals 2-P	Lognormals 3-P	Pearson 4-P	Pearson Gamma
Station 1	1960	109.8	43.4	43.2[a]	45.6	54.0
	1961	135.1	56.6[a]	62.6	109.3	110.1
	1962	129.9	45.0[a]	46.7	49.7	49.6
	1963	159.4	60.8[a]	93.1	131.6	210.6
	1964	119.5	59.3[a]	60.3	70.9	70.5
	1965	129.1	59.0	55.9	55.1[a]	84.8
	1966	129.0	52.2	52.4	48.3[a]	48.4
	1967	131.0	51.2[a]	52.5	60.6	63.9
	1968	120.3	60.2	59.4[a]	86.6	69.3
Average		129.2	54.2[a]	58.5	73.1	84.6
Station 2	1960	114.2	46.6	39.5	36.4[a]	39.1
	1961	125.7	41.6[a]	46.1	69.3	68.1
	1962	177.2	82.2	64.4[a]	105.7	68.2
	1963					
	1964	125.6	38.8[a]	39.2	62.3	53.4
	1965	119.5	43.5[a]	46.5	48.3	63.5
	1966	143.5	43.3	35.8[a]	52.1	67.0
	1967	143.7	59.8[a]	63.6	75.6	62.9
	1968	134.6	60.4[a]	61.3	80.9	68.9
Average		135.5	52.0	49.6[a]	66.3	61.4
Station 3	1960	108.6	47.3[a]	47.8	57.2	66.7
	1961	126.4	39.1[a]	39.3	56.9	53.1
	1962	106.0	27.7[a]	28.1	29.6	29.1
	1963	122.9	56.4	57.8	48.0	45.1[a]
	1964					
	1965	123.8	56.0	54.6[a]	57.4	89.4
	1966	134.9	58.6	66.8	50.0[a]	51.7
	1967	131.2	53.7	53.8	52.2	49.7[a]
	1968	136.9	61.1[a]	65.3	82.5	73.1
Average		123.8	50.0	51.7	54.2	57.2
Station 9	1968	94.0	60.3	58.9[a]	72.8	60.7
Station 11	1968	60.2	31.6[a]	36.0	36.9	32.0
Average		125.6	51.7[a]	53.0	64.1	66.8

[a]Best fit.

over the three-parameter lognormal, and the Pearson distributions fared considerably worse. However, note that in some cases each of the distributions provided the best fit. It is a curious fact that the

fitting method employed assigns a nonzero value to the location parameter despite the fact that zero will provide a better fit according to the sum of absolute deviations criterion. This provides a caveat, that this table might have been considerably different had a different fitting method or criterion been employed. This is seen from the fact that the 2-p lognormal fit better than the 3-p, of which it is a special case. Note too that the 2-p lognormal fared better than the 4-p distributions, a surprising fact due to the greater flexibility of the 4-p distributions.

Weibull Distribution

In 1951, Woloddi Weibull (25) published a paper in which the applicability of the distribution commonly written:

$$f(x) = Kx^m \exp\left[- Kx^{m+1}/(m+1)\right] \qquad (V\text{-}16)$$

was demonsirated. Although it had been known before 1951, the distribution has come to bear his name. The derivation presented therein is an interesting one since it is not derived from a single theoretical principle. Weibull approached the problem of finding the probability of failure of a chain consisting of n links P_n. He noted that the probability of nonfailure of the chain $1-P_n$ is equal to the probability of nonfailure of all the links simultaneously $(1-p)^n$ where p is the probability of failure of an individual link. Therefore, if each link has a distribution function governing its failure of the form

$$F(x) = 1 - e^{\psi(x)}, \qquad (V\text{-}17)$$

the distribution function for the chain will be

$$P_n = 1 - e^{-n\psi(x)} . \qquad (V\text{-}18)$$

The remaining problem is to specify $\psi(x)$. The only necessary condition is that it be a positive nondecreasing function, vanishing at μ which is not necessarily equal to 0. Weibull then stated that the simplest function satisfying this condition is

$$f(x) = \frac{(x - \mu)^m}{x_0} . \qquad (V\text{-}19)$$

The remarkable fact about this is that there is no theoretical justification for using this form, indeed Weibull states: "... it is utterly hopeless to expect a theoretical basis for distribution functions such as ... particle sizes."

As it happens, the Weibull distribution has been used to fit a large number of naturally occurring phenomena quite well. These include oil spill data, particle sizes, distances in cotton fibers, molecular weights, and solution concentrations.

The Weibull has both two- and three-parameter forms, the latter Eq. (V-16) being more common, where parameter m determines the shape of the curve and K is the scaling factor. Figure (11) illustrates the Weibull distributions for several values of m. Note that for m = 0 the Weibull reduces to the exponential, and for m = 1 it is equivalent to the Rayleigh distribution. It is obvious that when m = 1 or 2 the distributions appear similar to the lognormal. In fact, in cases where both distributions fit the same data, the Weibull shape parameter is always near this range.

The mean of the Weibull distribution is given by

$$E(x) = \left(\frac{K}{m+1}\right)^{1/m+1} \Gamma\left(\frac{m+2}{m+1}\right)\left(\frac{K}{m+1}\right)^{-2/m+1} \qquad \text{(V-20)}$$

and the variance is given by

$$\text{Var}(x) = \left[\Gamma\left(\frac{m+3}{m+1}\right) - \Gamma^2\left(\frac{m+2}{m+1}\right)\right] . \qquad \text{(V-21)}$$

As is usually the case with skew distributions in practical applications, the median is used as a measure of central tendency rather than the mean. The latter is extraordinarily sensitive to values in the tail of the skew distribution. This is true also for the lognormal.

For the Rayleigh,

$$E(x) = \sqrt{\frac{\pi}{2K}} \qquad \text{(V-22)}$$

$$\text{Var}(x) = \frac{2}{K}\left(1 - \frac{\pi^2}{4}\right) . \qquad \text{(V-23)}$$

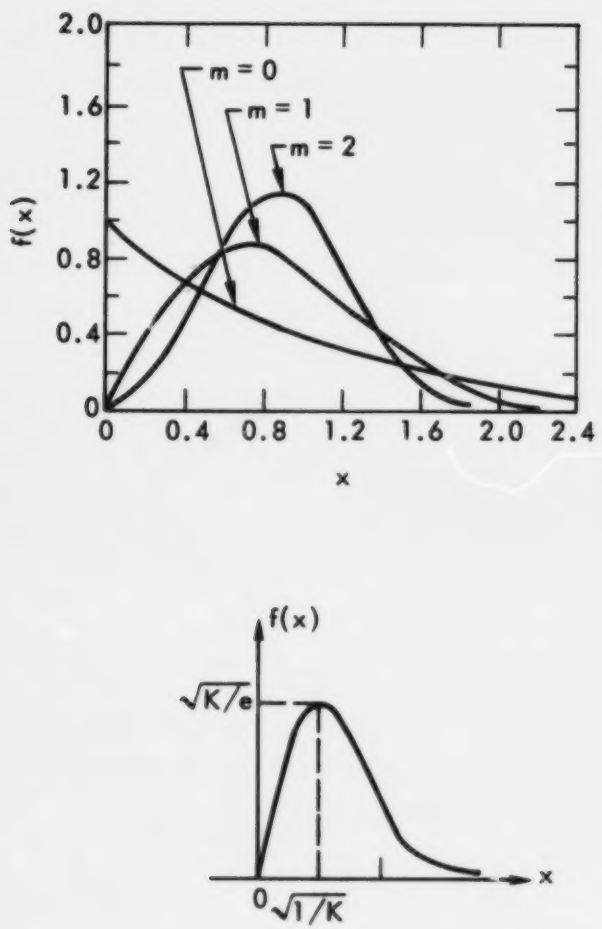

Fig. 11. Frequency curves for Weibull (top) and Rayleigh probability distributions.

It is interesting that the Weibull is usually fit with a narrow range for the shape parameter from one application to another. This suggests that the Rayleigh distribution may provide a fair fit, a surprising fact because it implies that a complex physical process is adequately described with only one parameter in a distribution with no theoretical foundation.

Figure 12 demonstrates the appearance of the Weibull probability distribution on log probability paper for parameter values typically obtained in pollution work.

Gamma Distribution

The gamma distribution is also used in air pollution work. We can easily see why from Fig. 13(a). The gamma has the ability to appear quite similar to both the lognormal and Weibull, depending on the values of the scale parameter β and the shape parameter α in

$$f(x) = \frac{1}{\beta^{\alpha+1}\,\Gamma(\alpha+1)}\, x^{\alpha}\, e^{-x/\beta} \qquad \text{(V-24)}$$

$$\alpha > -1,\ \beta > 0$$

$$0 < x \leq \infty .$$

The distribution can be derived as the distribution of the sum of n identical exponentially distributed random variables. The mean and variance of this distribution are given by

$$E(x) = \beta(\alpha + 1)$$

$$\text{Var}(x) = \beta^2(\alpha + 1) \qquad \text{(V-25)}$$

The gamma distribution also has a three-parameter form. The three-parameter form is derived by subtracting a location parameter from the mean, a process which also results in three-parameter forms for the other distributions. The three-parameter form is given by

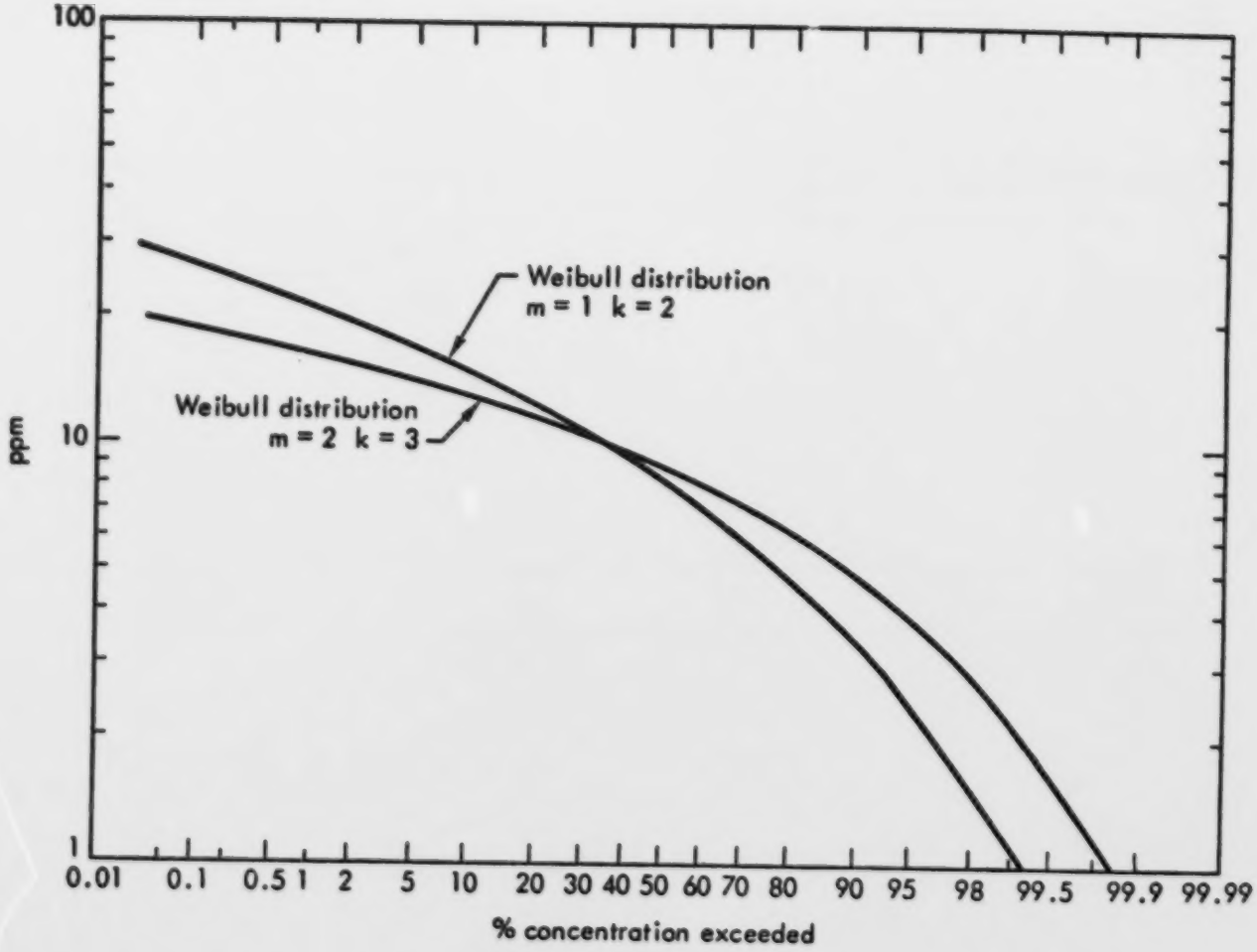
100
10
1
ppm
Weibull distribution
m = 1 k = 2
Weibull distribution
m = 2 k = 3
0.01 0.1 0.5 1 2 5 10 20 30 40 50 60 70 80 90 95 98 99.5 99.9 99.99
% concentration exceeded

Fig. 12. Cumulative Weibull distribution plotted on log probability paper.

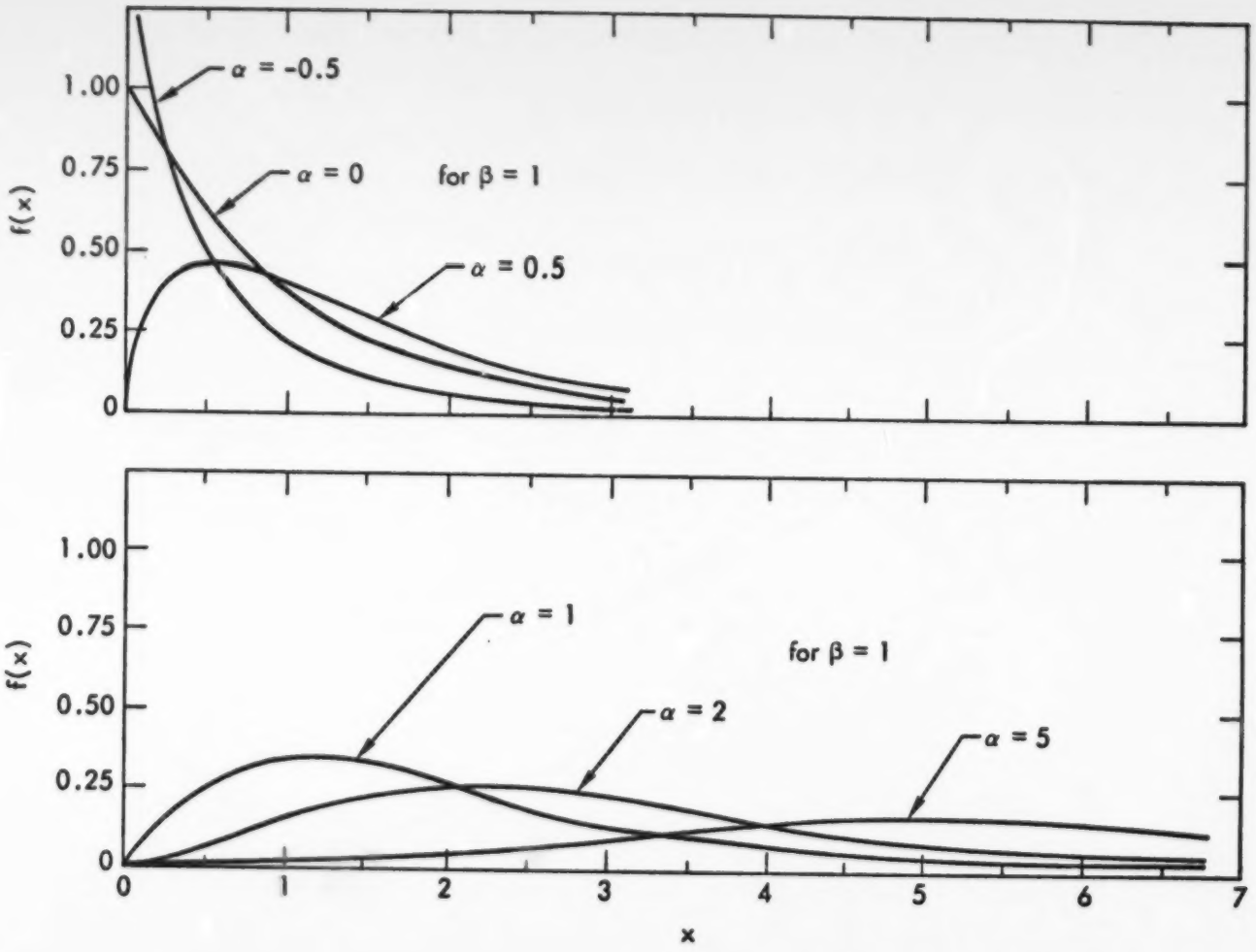

Fig. 13(a). Frequency curves for gamma probability distribution for various values of α.

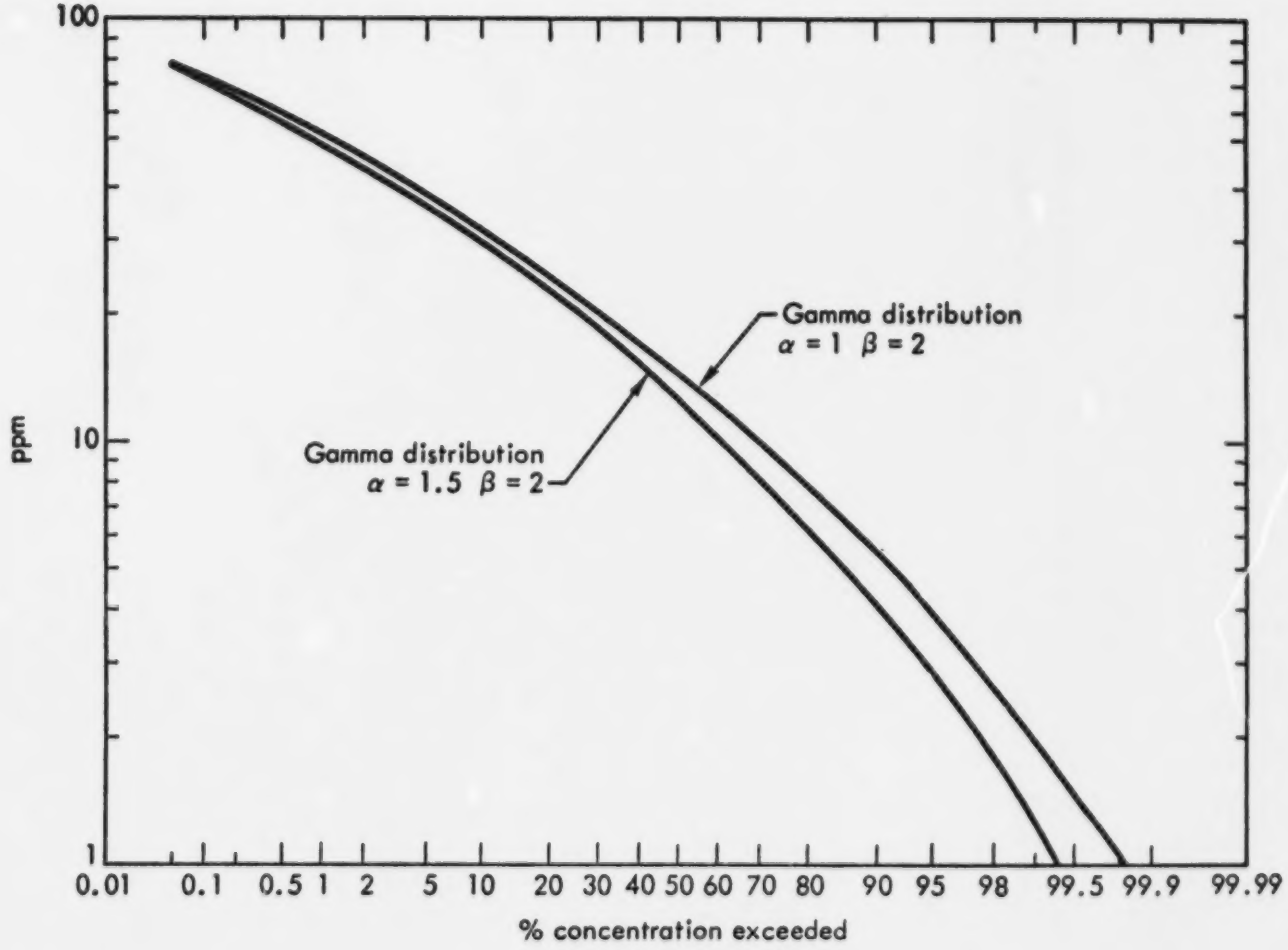
100
10
1
ppm
Gamma distribution
α = 1 β = 2
Gamma distribution
α = 1.5 β = 2
0.01 0.1 0.5 1 2 5 10 20 30 40 50 60 70 80 90 95 98 99.5 99.9 99.99
% concentration exceeded

Fig. 13(b). Cumulative gamma distributions plotted on log probability paper.

$$f(x) = \frac{1}{\beta^{\alpha+1}\Gamma(\alpha+1)} (x - x_0)^{\alpha} e^{-(x-x_0)/\beta} \qquad \text{(V-26)}$$

$$x > \gamma$$
$$\alpha > -1$$
$$\beta > 0 .$$

The gamma distribution is most widely used in reliability theory.

Figure 13(b) demonstrates that appearance of the Gamma probability distribution on log probability paper for parameter values typically obtained in pollution work.

Pearson Distribution

Pearson's system is to provide a theoretical density function for every possible combination of skewness and kurtosis (B_1, B_2) (see Fig. 14). There are three main types, I, IV, VI. Type I is the beta, type IV is the gamma. The procedure is to calculate B_1 and B_2 and see which part of the plane is indicated. For air quality data, type I often occurs. Type VI has been investigated also, although type IV was not needed.

The type I density is given by

$$f(x) = \frac{\Gamma(p) \cdot \Gamma(q)}{\Gamma(p+q)} \cdot \frac{(x - A)^{m_1} (B - x)^{m_2}}{(B - A)^{p+q-1}} \qquad \text{(V-27)}$$

4-p form,

$$m_1 = p - 1$$
$$m_2 = q - 1. \qquad \text{(V-28)}$$

The type VI is given by

$$y = \frac{(q_2 + 1)^{q_2} (q_1 - q_2 - 2)^{q_1 - q_2} \Gamma(q_1)}{x_0 (q_1 - 1)^{q_1} \Gamma(q_1 - q_2 - 1) \Gamma(q_2 + 1)} \left[\frac{\left(1 + \frac{x - \mu}{A_2}\right)^{q_2}}{\left(1 + \frac{x - \mu}{A_1}\right)^{q_1}} \right] \qquad \text{(V-29)}$$

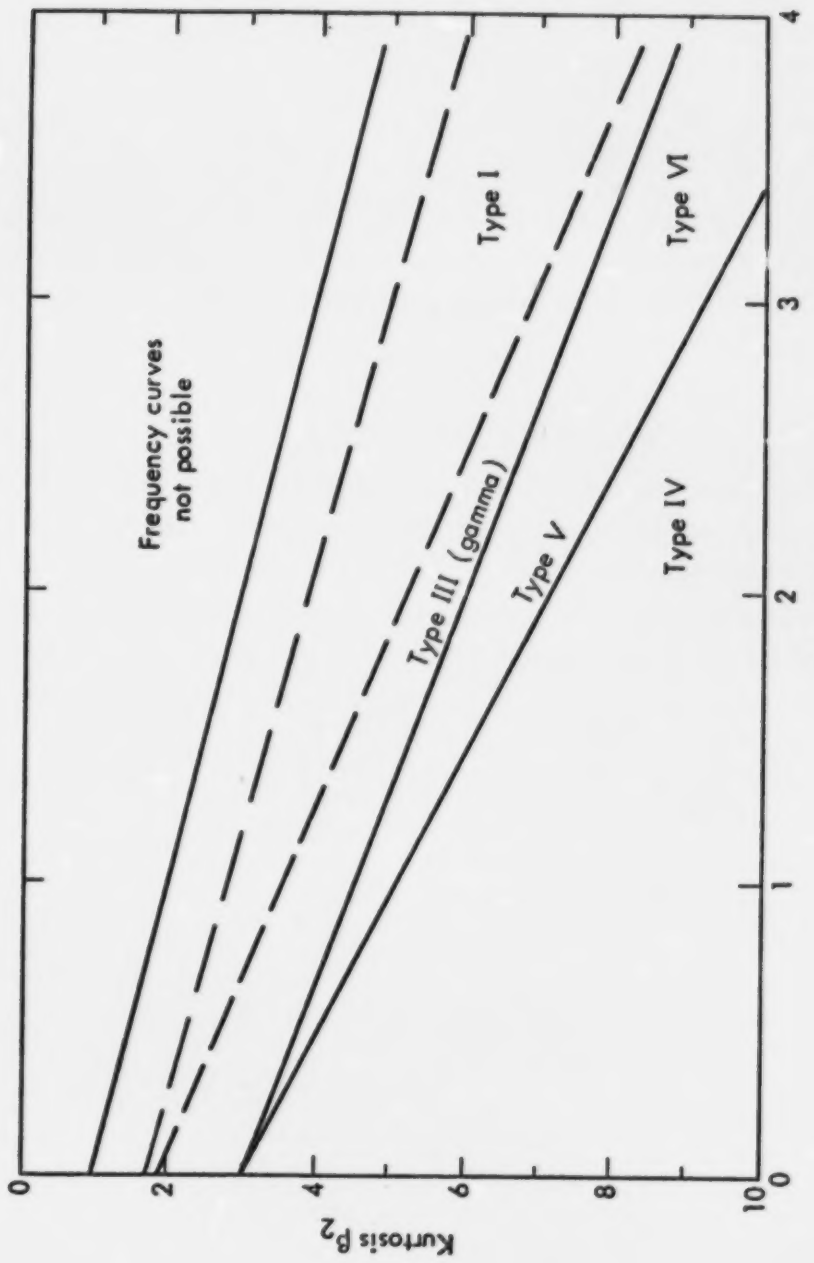

Fig. 14. Skewness-kurtosis plane in Pearson's system.

where

$$A_1 = \frac{x_0(q_1 - 1)}{(q_1 - 1) - (q_2 + 1)}$$

and

$$A_2 = \frac{x_0(q_2 + 1)}{(q_1 - 1) - (q_2 + 1)} \qquad \text{(V-30)}$$

where the origin is at zero ppm (or $\mu g/m^3$).

Figure 15 demonstrates the appearance of the beta probability distribution on log probability paper for parameter values typically obtained in pollution work.

Mathematical Similarity

A lingering question is how these distributions are similar mathematically. An exercise which gives considerable insight into this point is to try to find a transformation which reduces these distributions to similar form. One such transformation is to take the logarithm of the distribution and then differentiate it. For the lognormal,

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - a)^2}{2\sigma^2}} \qquad \text{(V-31)}$$

$$\ln[f(x)] = -\ln x - \ln\sigma - \ln\sqrt{2\pi} - \left[\frac{(\ln x)^2}{2\sigma^2} + \frac{a^2}{2\sigma^2} - \frac{2a\ln x}{2\sigma^2}\right] \qquad \text{(V-32)}$$

$$\frac{d[\ln(f(x))]}{dx} = -\frac{1}{x} - \frac{2\ln x}{2\sigma^2 x} + \frac{a}{\sigma^2 x} \qquad \text{(V-33)}$$

$$= \frac{1}{x}\left(-1 - \frac{a}{\sigma^2} - \frac{\ln x}{\sigma^2}\right). \qquad \text{(V-34)}$$

For the Weibull

$$f(x) = Kx^m \exp[-Kx^{m+1}/(m + 1)] \qquad \text{(V-35)}$$

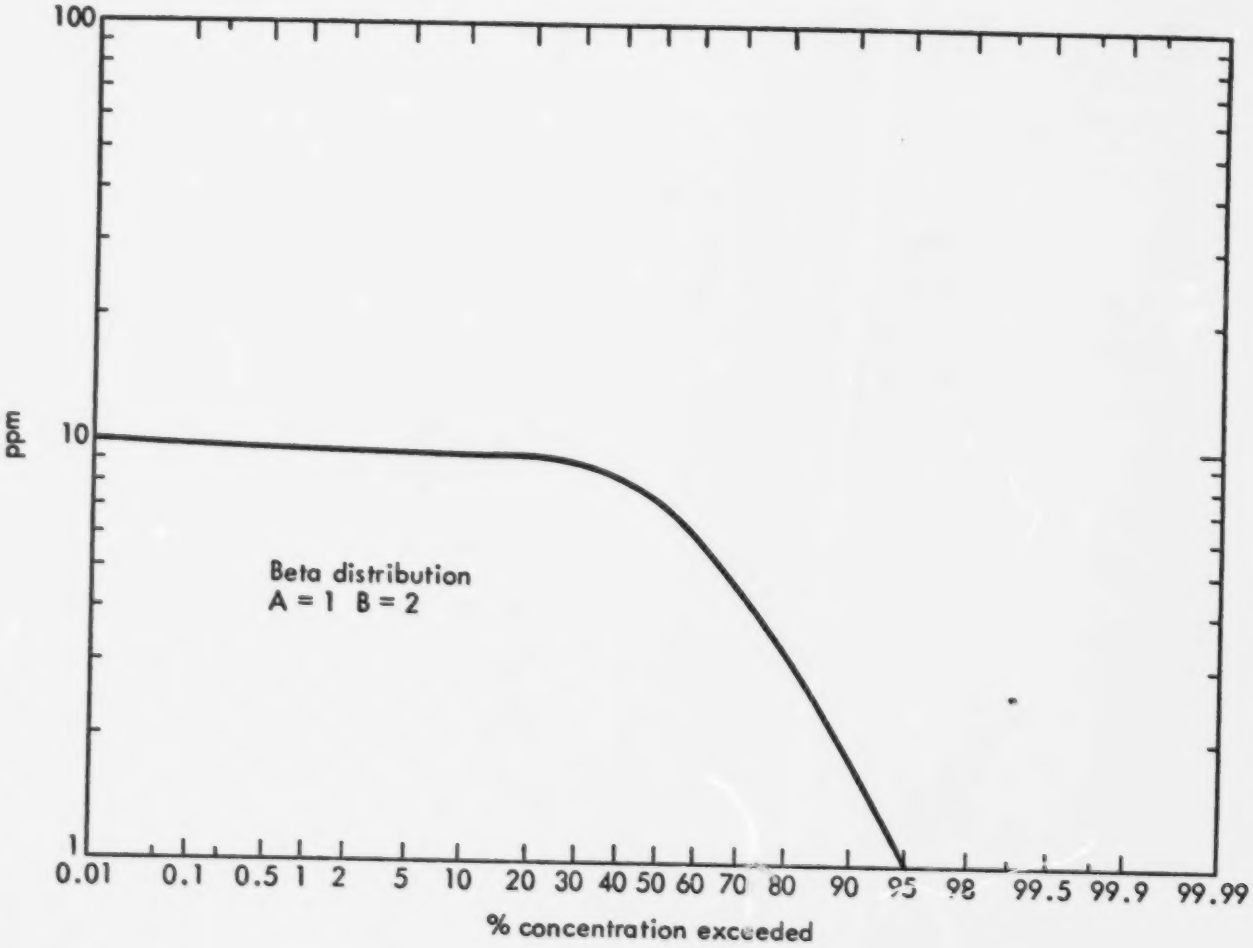

Fig. 15. Cumulative beta distribution plotted on log probability paper.

$$\ell n[f(x)] = \ln K + m \ln x - K x^{m+1}/(m+1) \tag{V-36}$$

$$\frac{d\{\ell n[f(x)]\}}{dx} = \frac{m}{x} - K x^{m} \tag{V-37}$$

$$= \frac{1}{x}(m - K x^{m+1}) . \tag{V-38}$$

For the gamma,

$$f(x) = \frac{1}{\alpha! \beta^{\alpha+1}} x^{\alpha} e^{-x/\beta} \tag{V-39}$$

$$\ell n[f(x)] = \ell n 1 - \ell n \alpha! - \ell n(\beta^{\alpha+1}) + \alpha \ell n x - x/\beta \tag{V-40}$$

$$\frac{d\{\ell n[f(x)]\}}{dx} = \frac{\alpha}{x} - \frac{1}{\beta} \tag{V-41}$$

$$= \frac{1}{x}(\alpha - x/\beta) . \tag{V-42}$$

Summarizing the final result for each:

$$\text{lognormal} = \frac{1}{x}\left(-1 + \frac{a}{\sigma^2} - \frac{\ell n x}{\sigma^2}\right)$$

$$\text{Weibull} = \frac{1}{x}(m - K x^{m+1})$$

$$\text{gamma} = \frac{1}{x}(\alpha - x/\beta) .$$

In each case there is a constant term, a function of the shape parameter, and a function of x. From experience in fitting those distributions, we know that the value of the appropriate parameters adjust so that the constant term is often between 1 and 2. The remaining term is higher order, and serves to provide the differences in goodness of fit noticed between these distributions.

Thus, the required theoretical results have been provided, with an additional section describing the similarity between these distributions.

Summary

In this chapter the air quality distributions presented earlier are discussed from a mathematical viewpoint. This serves to illustrate more clearly the nature of air quality distributions.

A result of significance to air pollution data analysis is the transformation which indicates a fundamental similarity between the various distributions used to fit air quality data. Future research may therefore be directed at modifying the second-order term to provide a more accurate distribution in cases where the two-parameter lognormal is inadequate because of the magnitude at the source term or the reactive nature of the pollutant.

CHAPTER VI—ILLUSTRATIVE APPLICATIONS

There is considerably greater utility in these results than the simple comparison of air quality to standards. Knowledge of the nature of these distributions and parameter variations under various conditions allow the construction of models for a variety of purposes. Several are outlined in the following section.

Analysis of Meteorological Patterns for Pollution Level Forecasting

A particularly useful application of these techniques is discussed below. It is included as an example of the power of the techniques presented earlier. Note that the fundamental philosophy of this application, i.e., that pollutant concentration distributions can be used as a partial substitute for meteorological data, is motivated by the arguments presented in this dissertation.

It has been established in Chapter II that the only variables affecting future pollutant concentrations are emissions, meteorological variables, photochemical change, and current concentration levels. It was further established that for all pollutants for all averaging times, concentrations are approximately lognormally distributed by both theoretical and empirical arguments. These assumptions can be used in the formulation of a predictive model. If one plots cumulative distribution functions on lognormal probability paper for individual days using hourly averages, the resulting curves are based on 24 points and, as expected from the theoretical argument in Chapter II, are relatively flat, straight lines. If one plots a great many such curves taken from data at one location over a certain period of time, it is possible that the resulting diagram will appear similar to Fig. 16. For regions where the climatology is very persistent, the clustering at the lines will be clear.

These lines can be mapped into points by plotting geometric mean (GM) versus SGD as in Fig. 17. In this plot, taken from actual Los Angeles (downtown) oxidant data for 1970, the degree of clustering is clear. It is also clear that changing the metric in which these points are plotted alters the degree of clustering seen from a plot. If we can find

Fig. 16. A possible set of air quality patterns.

the most independent clusters, with the total number of clusters constrained, we have then identified days with similar air quality patterns.

The significance of this is clear from Chapter II. Days with similar air quality patterns tend to have similar meteorological and emission patterns. Therefore, if we can identify the meteorology and emissions in each cluster and can then predict tomorrow's meteorological and emission pattern, we can determine in which cluster tomorrow's air quality will fall. Of course, each cluster refers to a particular GM and SGD which, as shown in Chapter II, completely describe a day's air quality from the point of view of air quality standards.

There is evidence (26) to the effect that emission patterns are primarily dependent on the day of week and time of year. Consequently, if we stratify our clustering graphs by day of week and season, we can then relate each cluster directly to a meteorological pattern.

The problems remaining are twofold:

1. How do we find the "best" clusters?
2. How do we determine into which cluster tomorrow's air quality falls?

Both of these problems could be handled pragmatically by an "eyeball" solution. In areas with high climatological persistence, this coarse method might give acceptable results. There are, however, exact methods to deal with these problems also.

If we refer to Fig. 17 and treat the points as nodes in a graph, we immediately realize that the clustering problem that has arisen from our air quality classification problem is mathematically equivalent to the clustering problem in modern graph theory. In fact, a great number of papers have been written concerning methods of solving for the optimal clusters. In general, the techniques do not propose to solve a large problem completely, but rather they deal with effective compromises which utilize the tradeoff between the distance the algorithm comes from the true optimum and the computing time necessary to reach that point.

In our case, the problem will not usually be very large. It is unlikely that more than three or four years of data would be analyzed together because of the gradual change in emissions. That leaves us with approximately 1400 points. As a working approximation, we can use only two significant figures which will serve to make many points identical.

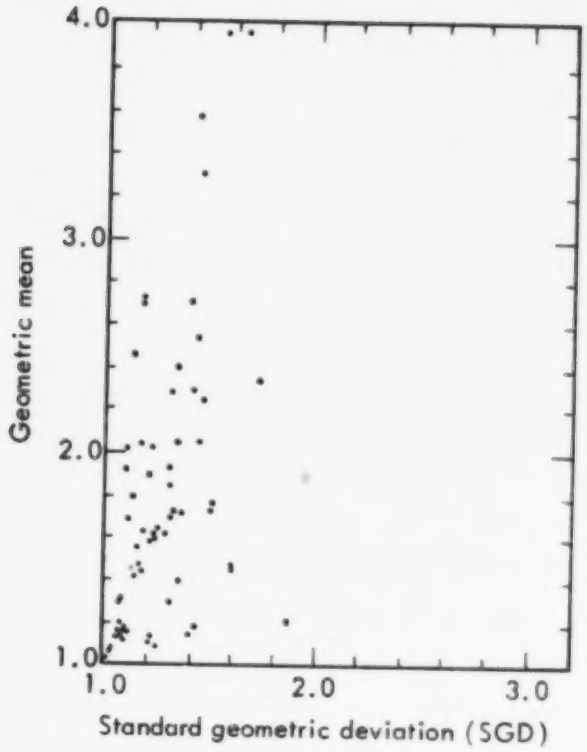

Fig. 17. Geometric mean versus standard geometric deviation for individual days for oxidant concentration in Los Angeles, California, 1970 hourly averages.

Hence we may expect to have about 10^3 points which we will try to divide into between 10 and 20 clusters. Several algorithms are available which will handle these numbers in a reasonable amount of computing time (27).

Selecting the Clusters

It is beyond the scope of this work to discuss the details of these algorithms; however, a simple explanation of a general method is in order. First, arbitrary clusters are selected and the bivariate median within each is calculated. Then, for each cluster, we calculate the sum of the distances from the bivariate median to each point within the cluster and the sum of the distances to each point out of the cluster. The objective function is to maximize the between-cluster differences minus the within-cluster differences. Each algorithm has a rule by which incremental changes in the clusters are made; at each stage the objective function is recalculated. Each algorithm also has a stopping rule based upon the size of the marginal improvement of a single or of a series of changes in the clusters.

Classifying New Days

The next problem to be dealt with arises after the final clustering arrangement has been identified. When the model goes into operation, the National Weather Service (NWS) forecasts are examined and the values of the predictor variables are selected. Now, from these values we must decide into which cluster the new day should be classified. This question should be answered even before the meteorological data are reduced.

The simplest method is to reduce data for the variable which are thought to be most important and construct a range of each variable for each category by "eyeball." Then the forecaster looks at the ranges thus selected and selects the cluster which the day most closely matches. A problem arises when the cluster is not distinct; in this case more variables or narrower ranges are needed. This information could also be displayed concisely in a series of nomographs, which would eliminate one source of error. A refinement of this technique would call for probability distributions of the ranges so that a value occurring near the

center of the range would be weighted more heavily than one near the extreme. Then point scores based upon probability could be used as the selection procedure.

A more rigorous method would be one that examines the probability distributions systematically and calculates both the classification and its probability of error. A well known statistical technique for which programs are available and which performs the required operations is multiple discriminant analysis (MDA).

In discriminant analysis, linear functions are developed which classify a new set of observations into one of several existing categories. The basic philosophy of the method is to define the categories to be used, in this case each cluster. Then, the values of the predictor variables associated with each point within each cluster are examined to discover patterns which will aid in the classification of a new set of predictors. A new metric for the predictors is found which maximizes the discrimination between the classes of predictors. Then based upon the within and between class distributions, conditional probability functions can be constructed which give the probability of membership in the ith class for a new set of predictor variables. A more complete discussion of the statistical techniques involved may be found in Ref. 28.

Recalibration

It is conceivable that over a period of several years the patterns of emissions and/or meteorology of an area will change in such a way as to affect the accuracy of the predictions made with this model. Fortunately, recalibration is accomplished in a relatively simple, straightforward manner.

Throughout the operation of the predictive model data should be kept, perhaps in a small notebook, noting the values of the variables used as predictors, the prediction, and the observed ambient air quality (AAQ). When the predicted and the observed values vary unacceptably the model is recalibrated, perhaps by the addition of a new category, either by an heuristic method or the "clustering" program, or by a full recalibration performed exactly as the original. Since the new data are available and the programs are already written, this procedure presents no problems. It is unlikely that it would be performed more often than biannually.

Spatial Interpolation

It should be clear that the foregoing analysis can predict concentration levels at only the receptor locations for which data sufficient for calibration are available. To predict air quality throughout a region, an interpolation scheme is required.

Because of the high correlation between wind velocities and concentrations of inert pollutants, especially over areas with relatively simple topography, interpolations can be made. It is necessary to make the assumption that concentration isopleths can be determined from the streamlines of the windfield, source location and available meteorological and concentration measurements (3). Given this and several receptors, one can identify the value of the isopleths passing through the receptors and interpolate for locations between the identified lines. Linear interpolation is adequate for areas which appear to be uniform in topography and emissions. Otherwise, experimental sampling, randomized spatially and temporally, may be employed to determine better interpolation functions. For inert pollutants from a point source, Benarie (3) has discussed this interpolation problem with respect to the frequency distributions. In his work the same fundamental assumption is made, but he makes another assumption which simplifies the calculation of the frequency distribution of an intermediate point. In particular, he assumes that the SGD, represented by the slope of the plot of the distribution function or lognormal probability paper, is constant throughout a streamline of the windfield. Therefore he requires only one point on the distribution function to estimate both parameters at an intermediate point along a windfield streamline for which he has the SGD calculated at another point. This method provides additional information with little effort and seems promising to be used in conjunction with the random sampling scheme mentioned above.

An Example

As an illustration of the techniques described herein, we shall present a simplified and to some extent hypothetical application of the model. Figure 17 is a graph of GM versus SGD for oxidant data in downtown Los Angeles in 1970. In Fig. 18 "eyeball" estimates of the clusters have been made.

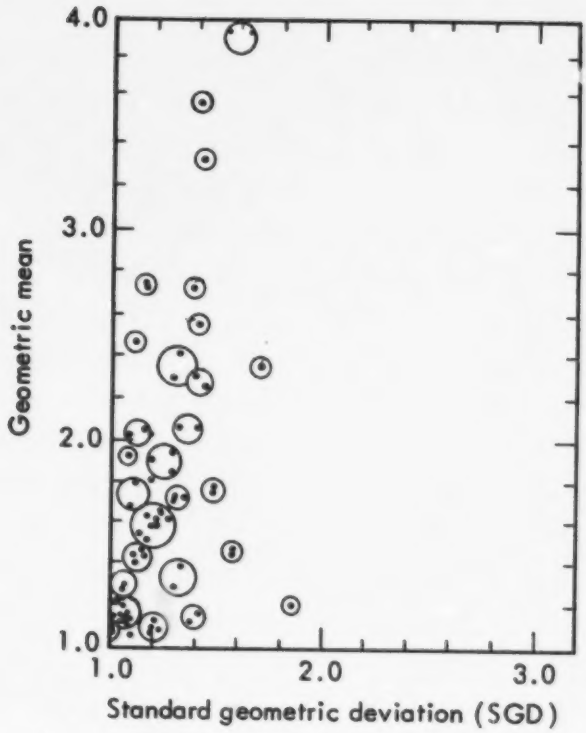

Fig. 18. A set of clusters of air quality day-types from the data in Fig. 17.

Now, according to the procedures outlined above, one takes a random sample of the points in each cluster and analyzes the meteorology for the day that point represents. This analysis requires a large amount of data reduction, and functions best with more than one year of data, stratified by season and day of week, to determine the values of windspeed, temperature and other variables used. Some of these data are recorded only in the archives of the NWS on synoptic scale maps which require a trained meteorologist to read. This is clearly beyond the scope of this work, despite the fact that this model is comparatively simple to calibrate. However, an actual run of the model is not essential for illustrative purposes. Instead we shall now turn to a hypothetical discussion of the kind to be expected in real application.

If we assume that we have successfully reduced data for a random sample of the days in each category, we can examine the range of values of the variables for each category. If we see, for example, that for the topmost cluster in Fig. 18 windspeeds are between 0 and 5 knots and the temperature varies between 90 and 95 deg, the oxidant concentrations are at episode levels. We record these values and proceed to the next cluster in each case, noting the mean, range and standard deviation of each variable.

Upon completion of this analysis we will be prepared to draw Fig. 19 which is a graph of the ranges of the two variables for each cluster and the identification of the cluster. Note that a nomographical technique would be required for a case with more than two predictors.

The shaded portion of the figure indicates an area of uncertainty where more than one cluster applies. The decision can be made using rigorous techniques such as discriminant analysis which gives both the classification and the probability of error, or by an heuristic technique such as determining how many standard deviations the center of the shaded portion is from the center of each of the two clusters and selecting the smaller. In effect, the latter method is a simplified "discriminant analysis."

Each cluster has its expected air quality level and a range of uncertainty, hence once the graph (Fig. 19) has been entered with the NWS predicted values, the problem is complete.

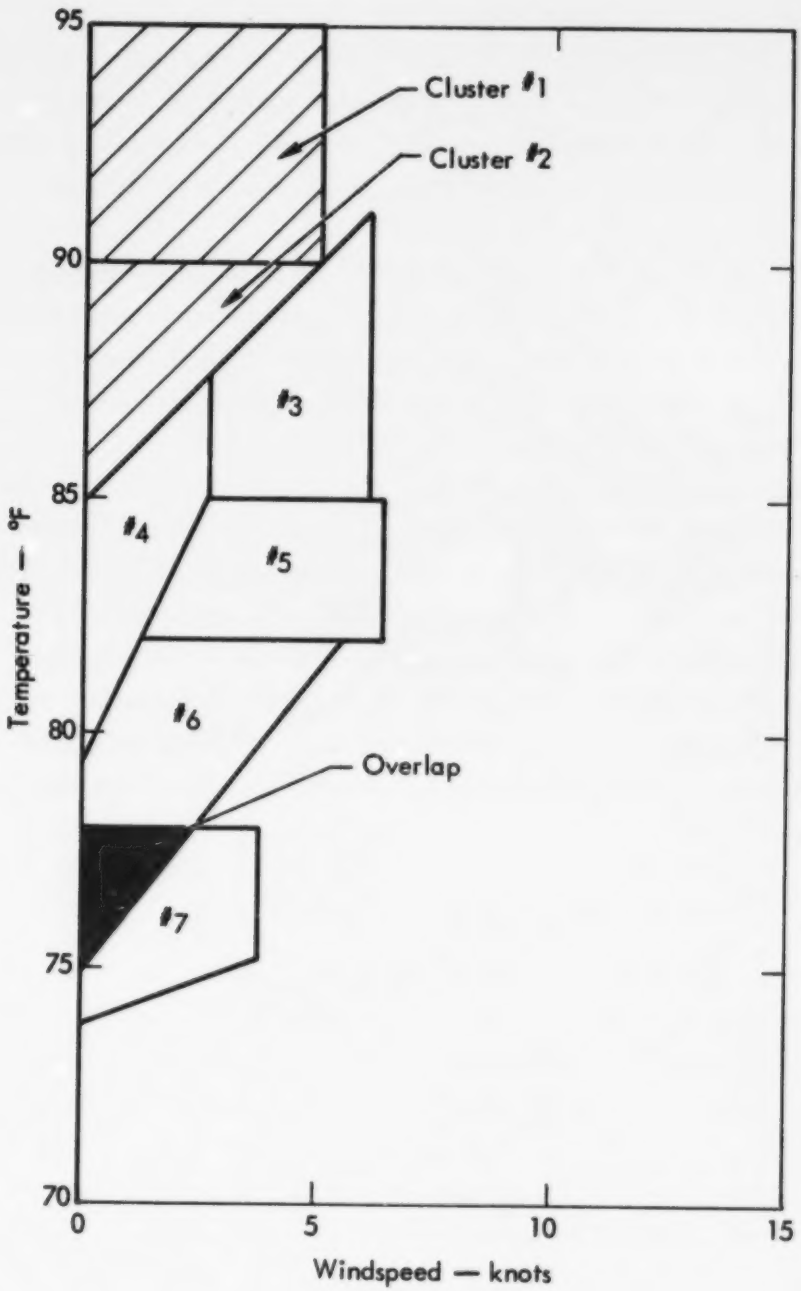

Fig. 19. An example of the form of the chart to be developed in comparing the clusters generated in Fig. 18 with windspeed and temperature.

Development

The model described above takes into consideration the expense of data collection and development work and the availability of computer time in that it minimizes the calibration data requirement, does not need a mesoscale weather prediction model because it adapts the standard NWS predictions, and does not require a computer on-line for predictions.

Further, the development effort required by the predictive model is also applicable to land use planning models and the comparison of present air quality with that of a benchmark year. For the latter application, the model eliminates bias in the comparisons due to differences in the meteorology of the years being compared.

This model was proposed to the California State Air Resources Board for use in the South Coast Air Basin to predict oxidant concentrations 24 hours in advance. The funding request was $80,000. Included in this amount was a full-time statistician, a part-time programmer and a full-time meteorologist.

Other modeling concepts are likely to be more expensive. For example, the multibox modeling concept requires considerably more effort and computer time, and least-squares analysis requires substantially more data.

Transition Matrices

A further application of the material presented herein involves the application of the categories defined above in land use planning. In particular, the air quality categories, stratified by emission-day-types, will be used to determine typical meteorological patterns, a problem previously tractable only subjectively by meteorologists examining large numbers of weather maps or by statisticians analyzing huge amounts of data much of which must be reduced from maps by trained meteorologists (29).

It can be argued that each of the clusters defined in the calibration of the predictive model represents a particular meteorological pattern. It is conceivable that two or more different patterns could constitute the same cluster; however, this is not necessarily significant because it is

unnecesary to distinguish between meteorological patterns yielding identical air quality for many purposes.

Therefore, if one calculates the frequency of occurrence of each pattern stratified by day of week and season, the results thus obtained allow one to simulate a year of "typical" meteorology which, in actuality, is a composite of all years for which air quality data exist. For most large cities, the continuous air monitoring program (CAMP) began in 1961. This "year," which is actually created as a Markov model using a state transition matrix, can be used in conjunction with a dispersion model as a benchmark year for comparing air quality over time, or in land use planning to determine future annual average pollutant concentrations.

Random Sampling

The justification of the lognormal assumption permits the use of parametric methods to operate on air quality data. An example of the usefulness of parametric methods as opposed to distribution-free methods is seen in the sample size required to estimate population parameters within a specified accuracy.

Table 4 indicates the sample sizes required under the various assumptions for oxidant data taken in Los Angeles in 1970. The efficiency of the parametric methods suggests that random sampling is an efficient method of characterizing regional air quality when an appropriate randomized scheme is used. Note that the measurements consisted of hourly averages every hour for an entire year. Clearly, a more cost effective scheme was possible.

Table 4. Sample size required for various confidence limits on estimates of GM and SGD.

PARAMETRIC		
Mean (Z-Statistic)		
At 95%	10%	90 samples
	5%	359 samples
At 99%	10%	155 samples
	5%	621 samples
Variance (χ^2 Statistic)		
At 95%	20%	200 samples
	10%	750 samples
	5%	5,000 samples
NONPARAMETRIC		
Mean (using Chebyshev Inequality)		
At 95%	10%	467 samples
	5%	1,869 samples
At 99%	10%	2,336 samples
	5%	9,344 samples
Variance [using Kolmogoroff's method (D-Statistic)]		
At 95%	10% on σ	3,000 samples (approx)
	5% on σ	18,500 samples (approx)

CHAPTER VII—SUMMARY AND CONCLUSIONS

The object of this work is to provide a model which binds together previous theoretical and empirical findings in a unified framework, and in so doing provides a deeper understanding of the physical processes which affect frequency distributions of air pollutant concentrations.

To this end, the frequency distributions of air pollutant concentrations have been derived from first principles for both point and area sources, for both reactive and inert pollutants. These results have been compared to published findings and have been found to be consistent.

Area Sources

Both Larsen's data analysis and Gifford and Hanna's simple model indicate that pollutant concentrations are approximately lognormally distributed. The former work consists of the examination of large quantities of data for all pollutants, for all cities and for all averaging times. The latter indicates the high degree of correlation between windspeeds, which are approximately lognormally distributed, and pollutant concentrations.

From the nonempirical standpoint, this distribution can be derived from the Fickian Diffusion equation by manipulating it into a finite difference form and demonstrating its consistency with the law of proportional effect. This method predicts the lognormal distribution will fit best for inert pollutants from area sources, and least well for reactive, secondary pollutants. Larsen's results and those of Gifford and Hanna are in good agreement with this assertion. This derivation also predicts that the lognormal distribution will not fit as well close to a source as it will further away. Recent data collected near large sources bear this out.

Also, a generalization of Gifford's point source model based upon the Gaussian Plume solution to the Fickian Diffusion equation indicates that pollutant concentrations are lognormally distributed if the geometric and arithmetic means are simply related.

Peripheral to the main discussion is an explanation of the surprising fact that pollutant concentrations are approximately lognormally distributed for all averaging times. This is explained through an analysis

of the averaging process as a window through which atmospheric motion of various scales can be seen.

As a result of this investigation, we are prepared to assert strongly that the lognormal is an appropriate distribution to use to characterize air quality data. We recognize that this will not significantly affect current practice, which has been proceeding on this basis, but will serve to quell the arguments concerning the correctness of this assumption, and lend further empirical and nonempirical support to those who are currently using the lognormal assumption. These users include parties responsible for monitoring air quality, meteorologists who are modeling atmospheric transport, and others who have use for these distributions along the lines suggested in Chapter VI.

Point Sources

No general agreement exists on the identity of the frequency distributions of air pollutants emanating from a point source. The empirical findings of Knox and Lange and Benarie are at odds with the theoretical prediction by Gifford. At present there is no explanation of these discrepancies in the literature.

In Chapter II a derivation is presented which indicates that any of the distributions mentioned in the literature may result depending on atmospheric stability, windspeeds, and the distance from source to receptor. Within this framework, each of the results in the literature may be obtained for the appropriate values of the relevant variables.

This suggests strongly that the eventual quantification of these relationships will proceed along the lines outlined here. This model is the first to reconcile the conflict through a treatment which provides understanding of the fundamental physical processes involved. It allows air pollution engineers to state with some certainty that pollutant concentration distributions resulting from a point source are neither lognormal nor chi-squared, but rather a subtle combination which depends upon the particular conditions under which the pollutant is measured.

Related Variables

The main point to be made here is that pollutant concentrations are tracers of atmospheric motion. As such, air quality frequency distribution data can be used as a partial substitute for meteorological data under certain conditions. An example is presented in Chapter VI.

To illustrate this point the cases of advection, diffusion and deposition are treated. Advective transport rates have been investigated empirically, and the lognormal distribution appears to fit quite well. Eddy diffusive transport rates can be demonstrated to be approximately lognormal by Kolmogorov's similarity theory argument, based upon energy exchange between different scales of turbulent motion. Deposition is based upon particle size distributions, which can be shown to be approximately lognormally distributed from both empirical and non-empirical arguments.

According to the "simple model" proposed and effectively applied by Gifford and Hanna, there is substantial correlation between windspeeds and pollutant concentrations. Based upon the lognormality of transport this statement is well motivated for nonreactive pollutants, especially in areas with relatively small source terms. In some cases the deposition component of the source term is also lognormally distributed, further contributing to the argument for the lognormality of pollutant concentrations.

Other Distributions

A number of authors have investigated the use of other frequency distributions to fit air quality data. The gamma, Weibull and beta distributions have received a good deal of attention. These distributions tend to fit marginally worse than the two-parameter lognormal, according to an extensive study by Lynn.

The fundamental questions are then: Why do these distributions do as well as indicated in the literature, and is there any theoretical support for these distributions to be used to characterize air quality data?

There are no reports in the literature presenting any nonempirical support for these distributions. However, the question remains that there must be some mathematical similarity between these distributions for the fits which have been observed to occur.

In this work a transformation has been found which transforms the lognormal, Weibull and gamma distributions to approximately the same form for typical parameter values observed in air quality data fits. This is indicative of a fundamental mathematical similarity between the distributions which demonstrates that if any one of the distributions fits the data the others must also, with only small differences in the goodness of fit.

This argument is significant in terms of the long standing discussion in the scientific community concerning which distribution is most appropriate, in that it gives a greater mathematical understanding of the goodness of fit observations. It also suggests a path for future research to determine the basic differences between the higher order terms of each distribution and how they relate to the physical processes at hand.

Applications

It is instructive to examine some of the modeling possibilities opened by the results presented here to demonstrate their utility. The applications outlined herein are actually being developed, or have been proposed for future development, in the author's work at the Lawrence Livermore Laboratory. It is expected that their utility will be demonstrated as the Laboratory effort progresses over the next several years.

Frequency distributions can be used to characterize meteorological and air quality patterns which have application in land use modeling and pollution level forecasting. They can also be used in air pollution dispersion modeling, and in the validation of such models.

The results presented herein and the accepted results in the literature justify these modeling concepts more firmly than the latter alone.

Future Research

For area sources a major question to be addressed is the prediction of the parameters of the concentration distribution. Research in this area is being conducted at present (30)(31).

The author has further work planned in studying the relationship of meteorological parameters to concentration distribution, particularly in the relationship of windspeeds to concentration.

The point source question is more complicated because of the changing identity of the distribution. The fundamental question here concerns the change in shape and parameters of the distribution as a function of windspeed, stability, reactivity of pollutant and distance. Perhaps other variables like stack height will also be significant. It will take a good deal more data than is currently available to produce definitive results on this matter.

The author has new work planned to delve more deeply into the matter of concentration distributions resulting from a point source. Simulation experiments are planned using the ADPIC (32) model to calculate such pollutant concentrations at various distances from the source under various meteorological regimes. The resulting frequency distributions will be compared with those predicted in Chapter II. This work should be completed in 1974.

The fundamental question concerning the identity of the distributions is not yet completely resolved. Additional theoretical and empirical support is still welcome, despite the strong arguments made in this work and in previous published reports.

LITERATURE CITED

1. R. I. Larsen and C. E. Zimmer, "Calculating Air Quality and Its Control," JAPCA, 15, 565 (1965).
   R. I. Larsen, "Analyzing Air Pollutant Concentration and Dosage Data," JAPCA, 17, 85 (1967).
   R. I. Larsen and C. E. Zimmer, "A New Mathematical Model of Air Pollutant Concentration Averaging Time and Frequency," JAPCA, 19, 24 (1969).
2. J. B. Knox and R. Lange, "Surface Air Pollutant Concentration Frequency Distribution: Implications for Urban Air Pollution Modelling," University of California, Lawrence Livermore Laboratory, Report UCRL-73887 (1972).
3. M. Benarie, "The Use of the Relationship Between Wind Velocity and Ambient Pollutant Concentration Distributions for the Estimation of Average Concentrations from Gross Meteorological Data," Proceedings of the Symposium on Statistical Aspects of Air Quality Data, Chapel Hill, North Carolina, November 1972.
   M. Benarie, "Sur La Validite De La Distribution Logarithmico-Normale Des Concentrations De Pollutant," Second International Clean Air Congress, 1970.
4. A. C. Stern, "Air Pollution," Vol. III, 2nd Ed. (Academic Press, New York, 1968).
5. F. Gifford, "Statistical Properties of a Fluctuating Plume Dispersion Model," Proceedings of the Symposium on Atmospheric Diffusion and Air Pollution, Oxford, August 1958.
6. F. Gifford, "The Form of the Frequency Distributions of Air Pollutant Concentrations," Proceedings of the Symposium on Statistical Aspects of Air Quality Data, Chapel Hill, North Carolina, November 1972.
7. J. B. Knox and R. I. Pollack, "An Investigation of the Frequency Distributions of Surface Air Pollutant Concentrations," Symposium on Statistical Aspects of Air Quality Data, Chapel Hill, North Carolina, November 1972.

8. M. C. MacCracken, T. V. Crawford, K. R. Peterson and J. B. Knox, "Initial Application of a Multi-Box Air Pollution Model to the San Francisco Bay Area," University of California, Lawrence Livermore Laboratory, Report UCRL-73994 (1972).
9. C. Hopper, personal communication, 1972.
10. N. A. Fuchs, The Mechanics of Aerosols, (Permagon Press, New York, 1964).
11. F. A. Gifford and S. R. Hanna, "Modeling Urban Air Pollution," ARATDL Contribution No. 63, 1972.
12. F. A. Gifford and S. R. Hanna, "Urban Air Pollution Modelling," presented at the 1970 International Air Pollution Conference of the International Union of Air Pollution Prevention Associations.
13. A. N. Kolomogoroff, Dokl. AN SSSR, 30, 301 (1941).
14. A. M. Yaglom, Dokl. AN SSSR, 166, 49 (1966).
15. A. S. Gurvich, Dokl. AN SSSR, 172, 554 (1967).
16. G. K. Batchelor, "The Application of the Similarity Theory of Turbulence to Atmospheric Diffusion," Quart. J. Roy. Met. Soc., 76, 133 (1950).
17. T. V. Crawford, "Atmospheric Diffusion of Large Clouds," Proceeding of the USAEC Meteorological Information Meeting, September 1967, Chalk River, Ontario Canada, Rept. AECL-2787 (1968).
18. T. V. Crawford, "A Computer Program for Calculating the Atmospheric Dispersion of Large Clouds," University of California, Lawrence Livermore Laboratory Report UCRL-50179 (1966).
19. I. H. Blifford and D. A. Gillette, "Applications of the Lognormal Frequency Distribution to the Chemical Composition and Size Distribution of Naturally Occurring Atmospheric Aerosols," Water, Air and Soil Pollution, 1, 106 (1971).
20. E. A. Shuck, J. N. Pitts, and J. K. S. Wan, "Relationships Between Certain Meteorological Factors and Photochemical Smog," Intern. J. Air Water Pollution, 10, 689 (1966).
21. R. L. Mitchell, "Permanence of the Lognormal Distribution," J. Opt. Soc. Am., 58, 1267 (1968).
22. D. S. Lynn, "Fitting Curves to Suspend Particulate Data," Proceedings of the Symposium on Statistical Aspects of Air Quality Data, Chapel Hill, North Carolina, November 1973.

23. P. G. Milokaj, "Environmental Applications of the Weibull Distribution Function: Oil Pollution," Science, 176 1019 (1972).
24. R. E. Barlow, "Averaging Time and Maxima for Air Pollution Concentration," NTIS AD-729 413, ORC 71-17.
25. W. Weibull, "A Distribution Function of Wide Applicability," J. Appl. Mech., 293 (1951).
26. E. Lawrence, "Urban Climate and Day of the Week," Atmos. Environ., 5, 935 (1971).
27. J. C. Gower, "A Comparison of Some Methods of Cluster Analysis," Biometrics, 23, 623 (1967).
28. R. G. Miller, "Statistical Predictions by Discriminant Analysis," Meteorol. Monographs, 4, 25 (1962).
29. C. L. Smalley, "A Survey of Air Flow Patterns in the San Francisco Bay Region 1952-1955," Bay Area Air Pollution Control District Technical Services Division Report.
30. R. Thullier, "Air Quality Statistics in Land Use Planning Applications," 3rd Conf. on Probability and Statistics in Atmospheric Science, Boulder, Colo., June 19-22, 1973.
31. W. B. Johnson, "The Status of Air Quality Simulation Modeling," Proceedings of the Interagency Conference on the Environment, Livermore, California, October 1972.
32. R. Lange, personal communication, 1973.